

Sophisticated Inference

Karl Friston

k.friston@ucl.ac.uk

Wellcome Centre for Human Neuroimaging, UCL Queen Square Institute of Neurology, London WC1N 3AR, U.K.

Lancelot Da Costa

l.da-costa@imperial.ac.uk

Wellcome Centre for Human Neuroimaging, UCL Queen Square Institute of Neurology, London WC1N 3AR, U.K., and Department of Mathematics, Imperial College London, U.K.

Danijar Hafner

mail@danijar.com

Department of Computer Science, University of Toronto, Toronto, ON M5S 2E4, Canada, and Google Research, Brain Team, Toronto, ON MSH 153, Canada

Casper Hesp

c.hesp@uva.nl

Wellcome Centre for Human Neuroimaging, UCL Queen Square Institute of Neurology, London WC1N 3AR, U.K., and Amsterdam Brain and Cognition Center, University of Amsterdam, Amsterdam 1001 NK, The Netherlands

Thomas Parr

thomas.parr.12@ucl.ac.uk

Wellcome Centre for Human Neuroimaging, UCL Queen Square Institute of Neurology, London WC1N 3AR, U.K.

Active inference offers a first principle account of sentient behavior, from which special and important cases—for example, reinforcement learning, active learning, Bayes optimal inference, Bayes optimal design—can be derived. Active inference finesses the exploitation-exploration dilemma in relation to prior preferences by placing information gain on the same footing as reward or value. In brief, active inference replaces value functions with functionals of (Bayesian) beliefs, in the form of an expected (variational) free energy. In this letter, we consider a sophisticated kind of active inference using a recursive form of expected free energy. Sophistication describes the degree to which an agent has beliefs about beliefs. We consider agents with beliefs about the counterfactual consequences of action for states of affairs *and* beliefs about those latent states.

In other words, we move from simply considering beliefs about “what would happen if I did that” to “what I would *believe about* what would happen if I did that.” The recursive form of the free energy functional effectively implements a deep tree search over actions and outcomes in the future. Crucially, this search is over sequences of belief states as opposed to states per se. We illustrate the competence of this scheme using numerical simulations of deep decision problems.

1 Introduction

In theoretical neurobiology, active inference has proved useful in providing a generic account of motivated behavior under ideal Bayesian assumptions, incorporating both epistemic and pragmatic value (Da Costa, Parr, Sajid et al., 2020; Friston, FitzGerald, Rigoli, Schwartenbeck, & Pezzulo, 2017). This account is often portrayed as being based on first principles because it inherits from the statistical physics of random dynamical systems at nonequilibrium steady state (Friston, 2013; Hesp, Ramstead et al., 2019; Parr, Da Costa, & Friston, 2020). Active inference does not pretend to replace existing formulations of sentient behavior; it just provides a Bayesian mechanics from which most (and, arguably, all) normative optimization schemes can be derived as special cases. Generally these special cases arise when ignoring one sort of uncertainty or another. For example, if we ignore uncertainty about (unobservable) hidden states that generate (observable) outcomes, active inference reduces to conventional schemes like optimal control theory and reinforcement learning. While the latter schemes tend to focus on the maximization of value as a function of hidden states per se, active inference optimizes a functional¹ of (Bayesian) beliefs about hidden states. This allows it to account for uncertainties surrounding action and perception in a unified, Bayes-optimal fashion.

Most current applications of active inference rest on the selection of policies (i.e., ordered sequences of actions or open-loop policies, where the sequence of future actions depends only on current states, not future states) that minimize a functional of beliefs called *expected free energy* (Da Costa, Parr, Sajid, et al., 2020; Friston, FitzGerald et al., 2017). This approach clearly has limitations, in the sense that one has to specify a priori allowable policies, each of which represents a possible path through a deep tree of action sequences. This formulation limits the scalability of the ensuing schemes because only a relatively small number of policies can be evaluated (Tschantz, Baltieri, Seth, & Buckley, 2019). In this letter, we consider active inference schemes that enable a deep tree search over all allowable sequences of

¹Technically, a functional is defined as a function whose arguments (in this case, beliefs about hidden states) are themselves functions of other arguments (in this case, observed outcomes generated by hidden states).

action into the future. Because this involves a recursive evaluation of expected free energy—and implicit Bayesian beliefs—the resulting scheme has a sophisticated aspect (Costa-Gomes, Crawford, & Broseta, 2001; Devaine, Hollard, & Daunizeau, 2014): rolling out beliefs about beliefs.

Sophistication is a term from the economics literature and refers to having beliefs about one's own or another's beliefs. For instance, in game theory, an agent is said to have a level of sophistication of 1 if she has beliefs about her opponent, 2 if she has beliefs about her opponent's beliefs about her strategy, and so forth. Most people have a level of sophistication greater than two (Camerer, Ho, & Chong, 2004).

According to this view, most current illustrations of active inference can be regarded as unsophisticated or naive, in the sense that they consider only beliefs about the consequences of action, as opposed to the consequences of action for beliefs. In what follows, we try to unpack this distinction intuitively and formally using mathematical and numerical analyses. We also take the opportunity to survey the repertoire of existing schemes that fall under the Bayesian mechanics of active inference, including expected utility theory (Von Neumann & Morgenstern, 1944), Bayesian decision theory (Berger, 2011), optimal Bayesian design (Lindley, 1956), reinforcement learning (Sutton & Barto, 1981), active learning (MacKay, 1992), risk-sensitive control (van den Broek, Wiegerinck, & Kappen, 2010), artificial curiosity (Schmidhuber, 2006), intrinsic motivation (Oudeyer & Kaplan, 2007), empowerment (Klyubin, Polani, & Nehaniv, 2005), and the information bottleneck method (Tishby, Pereira, & Bialek, 1999; Tishby & Polani, 2010).

Sophisticated inference recovers Bayes-adaptive reinforcement learning (Åström, 1965; Ghavamzadeh, Mannor, Pineau, & Tamar, 2016; Ross, Chaib-draa, & Pineau, 2008) in the zero temperature limit. Both approaches perform belief state planning, where the agent maximizes an objective function by taking into account how it expects its own beliefs to change in the future (Duff, 2002) and evinces a degree of sophistication. The key distinction is that Bayes-adaptive reinforcement learning considers arbitrary reward functions, while sophisticated active inference optimizes an expected free energy that can be motivated from first principles. While both can be specified for particular tasks, the expected free energy additionally mandates the agent to seek out information about the world (Friston, 2013, 2019) beyond what is necessary for solving a particular task (Tishby & Polani, 2010). This allows inference to account for artificial curiosity (Lindley, 1956; Oudeyer & Kaplan, 2007; Schmidhuber, 1991) that goes beyond reward seeking to the gathering of evidence for an agent's existence (i.e., its marginal likelihood). This is sometimes referred to as self-evidencing (Hohwy, 2016).

The basic distinction between sophisticated and unsophisticated inference was briefly introduced in appendix 6 of Friston, FitzGerald et al. (2017). As outlined in this appendix, there is a sense in which unsophisticated formulations, which simply sum the expected free energy over future time

steps based on current beliefs about the future, can be thought of as selecting policies that optimize a path integral of the expected free energy. In contrast, sophisticated schemes take account of the way in which the free energy changes as alternative paths are pursued and beliefs updated. This can be thought of as an expected path integral.

This distinction is subtle but can lead to fundamentally different kinds of behavior. A simple example illustrates the difference. Consider the following three-armed bandit problem—with a twist. The right and left arms increase or decrease your winnings. However, you do not know which arm is which. The central arm does not affect your winnings but tells you which arm pays off. Crucially, once you have committed to either the right or the left arm, you cannot switch to the other arm. This game is engineered to confound agents whose choice behavior is based on Bayesian decision theory. This follows because the expected payoff is the same for every sequence of moves. In other words, choosing the right or left arm—for the first and subsequent trials—means you are equally likely to win or lose. Similarly, choosing the middle arm (or indeed doing nothing) has the same Bayesian risk or expected utility.

However, an active inference agent, who is trying to minimize her expected free energy,² will select actions that minimize the risk of losing and resolve her uncertainty about whether the right or left arm pays off. This means that the center arm acquires epistemic (uncertainty-resolving) affordance and becomes intrinsically attractive. On choosing the central arm—and discovering which arm holds the reward—her subsequent choices are informed, in the sense that she can exploit her knowledge and commit to the rewarding arm. In this example, the agent has resolved a simple exploration-exploitation dilemma³ by resolving ambiguity as a prelude to exploiting updated beliefs about the consequences of subsequent action. Note that because the central arm has been selected, there is no ambiguity in play, and its epistemic affordance disappears. Note further that all three arms initially have some epistemic affordance; however, the right and left arms are less informative if the payoff is probabilistic.

²Expected free energy can be read as risk plus ambiguity: *risk* is taken here to be the relative entropy (i.e., KL divergence) between predicted and preferred outcomes, while *ambiguity* is the conditional entropy (i.e., conditional uncertainty) about outcomes given their causes.

³Exploration here has been associated with the resolution of ambiguity or uncertainty about hidden states, namely, the context in which the agent is operating (i.e., left or right arm payoff). More conventional formulations of exploration could remove the prior belief that the right and left arms have a complementary payoff structure, such that the agent has to learn the probabilities of winning and losing when selecting either arm. However, exactly the same principles apply: the right and left arms now acquire an epistemic affordance in virtue of resolving uncertainty about the contingencies that underlie payoffs as opposed to hidden states. We will see how this falls out of expected free energy minimization later.

The key move behind this letter is to consider a sophisticated agent who evaluates the expected free energy of each move recursively. Simply choosing the central arm to resolve uncertainty does not, in and of itself, mean an epistemic action was chosen in the service of securing future rewards. In other words, the central arm is selected because all the options had the same Bayesian risk⁴ while the central arm had the greatest epistemic affordance.⁵ Now consider a sophisticated agent who imagines what she will do after acting. For each plausible outcome, she can work out how her beliefs about hidden states will be updated and evaluate the expected free energy of the subsequent move under each action and subsequent outcome. By taking the average over both, she can evaluate the expected free energy of the second move that is afforded by the first. If she repeats this process recursively, she can effectively perform a deep tree search over all ordered sequences of actions and their consequences.

Heuristically, the unsophisticated agent simply chooses the central arm because she knows it will resolve uncertainty about states of affairs. Conversely, the sophisticated agent follows through—on this resolution of ambiguity—in terms of its implications for subsequent choices. In this instance, she knows that only two things can happen if she chooses the central arm: either the right or left arm will be disclosed as the payoff arm. In either case, the subsequent choice can be made unambiguously to minimize risk and secure her reward. The average expected free energy of these subsequent actions will be pleasingly low, making a choice of the central arm more attractive than its expected free energy would otherwise suggest. This means the sophisticated agent is more confident about her choices because she has gone beyond forming beliefs about the consequences of action to consider the effects of action on subsequent beliefs and the (epistemic) actions that ensue. The remainder of this letter unpacks this recursive kind of planning, using formal analysis and simulations.

This letter is intended to introduce a sophisticated scheme for active inference and provide some intuition as to how it works in practice. We validate this scheme through reproducing simulation results from previous formulations of active inference in a simple and a more complex navigation task. This is not intended as proof of superiority of sophisticated inference over existing schemes, which we assess in a companion paper (Da Costa, Sajid et al. 2020), but to demonstrate noninferiority in some illustrative settings. Note that it is possible to show that on reward maximization tasks, sophisticated active inference will significantly outperform, as it accommodates the backward induction algorithm as a special case.

⁴ Bayesian risk is taken to be negative expected utility, that is, expected loss under some predictive posterior beliefs (about hidden states).

⁵ Epistemic affordance is taken to be the information gain or relative entropy of predictive beliefs (about hidden states) before and after an action.

This paper has four sections. Section 2 provides a brief overview of active inference in terms of free energy minimization and the various schemes that can be used for implementation. This section starts with the basic imperative to optimize Bayesian beliefs about latent or hidden states of the world in terms of approximate Bayesian (i.e., variational) inference (Dayan, Hinton, Neal, & Zemel, 1995). It then goes on to cast planning as inference (Attias, 2003; Botvinick & Toussaint, 2012) as the minimization of an expected free energy under allowable sequences of actions or policies (Friston, FitzGerald et al., 2017). The foundations of expected free energy are detailed in an appendix from two complementary perspectives, the second of which is probably more fundamental as it rests on the first-principle account mentioned above (Friston, 2013, 2019; Parr et al., 2020). The third section considers sophisticated schemes using a recursive formulation of expected free energy. Effectively, this enables the efficient search of deep policy trees (that entail all possible outcomes under each policy or path). This search is efficient because only paths that have a sufficiently high predictive posterior probability need to be evaluated. This restricted tree search is straightforward to implement in the present setting because we are propagating beliefs (i.e., probabilities) as opposed to value functions. The fourth section provides some illustrative simulations that compare sophisticated and unsophisticated agents in the three-armed bandit (or T-maze paradigm) described above. It also considers deeper problems, using navigation and novelty seeking as an example. We conclude with a brief summary of what sophisticated inference brings to the table.

2 Active Inference and Free Energy Minimization

Most of the active inference literature concerns itself with partially observable Markov decision processes. In other words, it considers generative models of discrete hidden states and observable outcomes, with uncertainty about the (likelihood) mapping between hidden states and outcomes and (prior) probability transitions among hidden states. Crucially, sequential policy selection is cast as an inference problem by treating sequences of actions (i.e., policies) as random variables. Planning then simply entails optimizing posterior beliefs about the policies being pursued and selecting an action from the most likely policy.

On this view, there are just two sets of unknown variables: hidden states and policies. Belief distributions over this bipartition can then be optimized with respect to an evidence bound in the usual way, using an appropriate mean-field approximation (Beal, 2003; Winn & Bishop, 2005). In this setup, we can associate *perception* with the optimization of posterior beliefs about *hidden states*, while *action* follows from planning based on posterior beliefs about *policies*. Implicit in this formulation is a generative model: a probabilistic specification of the joint probability distribution over policies, hidden states, and outcomes. This generative model is usually factorized into

the likelihood of outcomes, given hidden states, the conditional distribution over hidden states, given policies, and priors over policies. In active inference, the priors over policies are determined by their expected free energy, noting that this energy, which depends on future courses of action, furnishes an empirical prior over subsequent actions.

In brief, given some prior beliefs about the initial and final states of some epoch of active inference, the game is to find a posterior belief distribution over policies that brings the initial distribution as close as possible to the final distribution, given observations. This objective can be achieved by optimizing posterior beliefs about hidden states and policies with respect to a variational bound on (the logarithm of) the marginal likelihood of the generative model (i.e., log evidence). This evidence bound is known as a variational free energy or (negative) evidence lower bound. In what follows, we offer an overview of the formal aspects of this enactive kind of inference.

2.1 Discrete State-Space Models. Our objective is to optimize beliefs (i.e., an approximate posterior) over policies π and their consequences, namely, hidden states $s \equiv s_{\leq \tau}$ from some initial state s_1 , until some policy horizon τ , given some observations $o_{\leq t}$ up until the current time t . This optimization can be cast as minimizing a (generalized) free energy functional $F[Q(s, \pi)]$ of the approximate posterior (Parr & Friston, 2019b). This generalized free energy has two parts: a generative model for state transitions, given policies, and a generative model for policies that depend on the final states (omitting constants for clarity):

$$\begin{aligned}
 F[Q(s, \pi)] &= \mathbb{E}_{Q(\pi)}[F(\pi)] + D_{KL}[Q(\pi) || P(\pi)] \\
 &= \mathbb{E}_{Q(\pi)}[\ln Q(\pi) + E(\pi) + F(\pi) + G(\pi)] \\
 F(\pi) &= \mathbb{E}_{Q(s_{\leq \tau} | \pi)}[\ln Q(s_{\leq \tau} | \pi) - \ln P(o_{\leq t}, s_{\leq \tau} | \pi)] \\
 G(\pi) &= \mathbb{E}_{Q(o_{\tau}, s_{\tau} | \pi)}[\ln Q(s_{\tau} | \pi) - \ln P(o_{\tau}, s_{\tau})] \\
 Q(o_{\tau}, s_{\tau} | \pi) &= P(o_{\tau} | s_{\tau}) Q(s_{\tau} | \pi) \\
 -\ln P(\pi) &= E(\pi) + G(\pi) \tag{2.1}
 \end{aligned}$$

This generalized free energy includes the variational free energy⁶ of each policy $F(\pi)$ that depends on priors over state transitions and an expected free energy of each policy $G(\pi)$ that underwrites priors over policies. The priors over policies $\ln P(\pi) = -E(\pi) - G(\pi)$ ensure the expected free energy at time τ (i.e., the policy horizon) is minimized. Here, $E(\pi)$ represents an empirical prior that is usually conditioned on hidden states at a higher level in deep (i.e., hierarchical) generative models. Note that outcomes on

⁶Note that both $F[Q(s, \pi)]$ and $F(\pi)$ depend on present and past observations. However, this dependence is typically left implicit, a convention we adhere to in this letter.

the horizon are random variables with a likelihood distribution, whereas outcomes in the past are realized variables. The distributions indicated by Q are variational distributions that have various interpretations throughout this letter. They inherit these interpretations in virtue of when we are in time. This means they are posterior probabilities when we account for data that have already been observed but can play the role of (empirical) priors when thinking about observations that have yet to be observed.

The first equality shows that the variational free energy, expected under the posterior over policies, plays the role of an accuracy, while the complexity of posterior beliefs about policies is the divergence from prior beliefs.⁷ In other words, variational free energy scores the evidence for a particular policy that accrues from observed outcomes. The priors over policies also have the form of a free energy. For interested readers, the appendix provides a fairly comprehensive motivation of this functional form, from complementary perspectives. In addition, Table 1 provides a glossary of variables used in this letter. We now consider the role of free energy in exact, approximate, and amortized inference, before turning to active inference and policy selection.

2.2 Perception as Inference. Optimizing the posterior over hidden states renders the variational free energy equivalent to (negative) log evidence—or marginal likelihood—in the usual way while optimizing the posterior over policies renders the generalized free energy zero:

$$\begin{aligned}
 Q(s|\pi) &= \arg \min_Q F(\pi) = P(s|o_{\leq t}, \pi) \\
 &\Rightarrow F(\pi) = -\ln P(o_{\leq t}|\pi) \\
 Q(\pi) &= \arg \min_Q F[Q(s, \pi)] \\
 &= \sigma[-E(\pi) - F(\pi) - G(\pi)] \\
 &\Rightarrow F[Q(s, \pi)] = 0
 \end{aligned} \tag{2.2}$$

The first equalities correspond to exact Bayesian inference based on a softmax function (i.e., normalized exponential, $\sigma[\cdot]$) of the log probability over outcomes and hidden states, under a particular policy. To finesse the numerics of optimizing the posterior over all hidden states, a mean-field approximation usually leverages the Markovian form of the generative model to optimize an approximate posterior over hidden states at each time point

⁷Generally log evidence is accuracy minus complexity, where accuracy is the expected log likelihood and complexity is the KL divergence between posterior and prior beliefs.

Table 1: Glossary of Variables.

Notation	Variable
$P(\cdot)$	Probability distribution
$Q(\cdot)$	Variational posterior or empirical prior distribution
F	Variational free energy
G	Expected free energy
u_τ	Action at time τ
$o = (o_1, o_2, \dots, o_\tau, \dots)$	Observation
$s = (s_1, s_2, \dots, s_\tau, \dots)$	Hidden (latent) states
π	Policy (sequence of actions)
s_τ^π	Expectation of state at time τ under $Q(s_\tau \pi)$
s_τ^u	Expectation of state at time τ under $Q(s_\tau u_\tau)$
v_τ^π	Log expectation of state at time τ under $Q(s_\tau \pi)$
o_τ^u	Expectation of observation at time τ under $Q(o_\tau u_{<\tau})$
u_τ^o	Expectation of action at time τ under $Q(u_\tau o_\tau)$
\mathbf{A}	Parameters of categorical likelihood distribution
\mathbf{B}	Parameters of categorical transition probabilities
\mathbf{C}	Parameters of categorical prior preferences
\mathbf{D}	Parameters of categorical initial state probabilities
\mathbf{H}	Conditional entropy of likelihood distribution
a, \mathbf{a}	Prior and posterior Dirichlet parameters for \mathbf{A}
b, \mathbf{b}	Prior and posterior Dirichlet parameters for \mathbf{B}
d, \mathbf{d}	Prior and posterior Dirichlet parameters for \mathbf{D}
$Cat(\cdot)$	Categorical probability distribution
$Dir(\cdot)$	Dirichlet probability distribution
$\mathbb{E}_P[\cdot]$	Expectation under the subscripted probability distribution
$H[\cdot]$	Shannon entropy of a probability distribution
$D_{KL}[\cdot\ \cdot]$	Kullback-Leibler divergence between probability distributions
$\psi(\cdot)$	Digamma function
$\sigma(\cdot)$	Softmax (normalized exponential) function

(where $s_{\setminus\tau}$ denotes the Markov blanket of s_τ):

$$\begin{aligned}
 Q(s_\tau|\pi) &= \sigma[\mathbb{E}_{Q(s_{\setminus\tau}|\pi)}[\ln P(o_{\leq t}, s_{\leq t}|\pi)]] \\
 &= \sigma[\mathbb{E}_{Q(s_{\setminus\tau}|\pi)}[\ln P(o_\tau|s_\tau) + \ln P(s_\tau|s_{\tau-1}, \pi) + \ln P(s_{\tau+1}|s_\tau, \pi)]] \\
 Q(s|\pi) &= Q(s_1|\pi)Q(s_2|\pi) \dots Q(s_\tau|\pi) \\
 P(s|\pi) &= P(s_1|\pi)P(s_2|s_1, \pi) \dots P(s_\tau|s_{\tau-1}, \pi)
 \end{aligned} \tag{2.3}$$

This corresponds to a form of approximate Bayesian inference (i.e., variational Bayes) in which equation 2.3 is iterated over the factors of the mean-field approximation to perform a coordinate descent or fixed-point iteration (Beal, 2003). An alternative formulation rests on an explicit minimization of variational free energy using iterated gradient flows to each fixed point

(expressed in terms of sufficient statistics):

$$\begin{aligned}\dot{\mathbf{v}}_{\tau}^{\pi} &= -\partial_{\mathbf{s}_{\tau}^{\pi}} F(\pi) = \mathbb{E}_{Q(\mathbf{s}_{\leq \tau}|\pi)}[\ln P(o_{\leq \tau}, \mathbf{s}_{\leq \tau}|\pi)] - \ln Q(\mathbf{s}_{\tau}|\pi) \\ \mathbf{s}_{\tau}^{\pi} &= \sigma(\mathbf{v}_{\tau}^{\pi}) \\ Q(\mathbf{s}_{\tau}|\pi) &= \text{Cat}(\mathbf{s}_{\tau}^{\pi})\end{aligned}\tag{2.4}$$

This solution can be read as (neuronal) dynamics that implement variational message passing⁸ (Beal, 2003; Friston, Parr, & de Vries, 2017; Parr, Markovic, Kiebel, & Friston, 2019). In this form, the free energy gradients constitute a prediction error: the difference between the posterior surprisal⁹ and its predicted value.

Finally, one can consider amortizing inference using standard procedures from machine learning to optimize the parameters ϕ of a recognition model with regard to variational free energy. In the present setting, this approach can be summarized as using universal function approximators (e.g., deep neural networks) to parameterize equation 2.2, namely, the mapping between observations and the sufficient statistics of the approximate posterior—for example,

$$\begin{aligned}\mathbf{s}_{\tau}^{\pi} &= f_{\phi}(o_{\leq \tau}, \mathbf{s}_{\leq \tau}^{\pi}, \pi) \\ \phi &= \arg \min_{\phi} F[Q(\mathbf{s}, \pi)] \\ Q(\mathbf{s}_{\tau}|\pi) &= \text{Cat}(f_{\phi})\end{aligned}\tag{2.5}$$

Effectively, amortized inference is “learning to infer” (Çatal, Nauta, Verbeelen, Simoens, & Dhoedt, 2019; Lee & Keramati, 2017; Millidge, 2019; Toussaint & Storkey, 2006; Tschantz et al., 2019; Ueltzhöffer, 2018). Variational autoencoders can be regarded as an instance of amortized inference, if we ignore conditioning on policies (Suh, Chae, Kang, & Choi, 2016). Clearly, amortization precludes online inference and may appear biologically implausible. However, it might be the case that certain brain structures learn to infer; for example, the cerebellum might learn from inferential processes implemented by the cerebral cortex (Doya, 1999; Ramnani, 2014).

2.3 Planning as Inference. The posterior over policies is somewhat simpler to evaluate—as a softmax function of their empirical,¹⁰ variational, and

⁸Where v can be thought of as transmembrane voltage or depolarization and s corresponds to the average firing rate of a neuronal population. (Da Costa, Parr, Sengupta, & Friston, 2020).

⁹Surprisal is the self-information or negative log probability of outcomes. (Tribus, 1961).

¹⁰The empirical free energy is usually based on inferences at a higher level in a hierarchical generative model. For details on hierarchical generative models, see Friston, Rosch, Parr, Price, and Bowman (2017).

expected free energy. This can be expressed in terms of a generalized free energy that includes the parameters of the generative model (e.g., the likelihood parameters, A):

$$\begin{aligned}
 Q(\pi) &= \arg \min_Q F[Q(s, \pi, A)] = \sigma[-E(\pi) - F(\pi) - G(\pi)] \\
 G(\pi) &= \mathbb{E}_{Q(o_\tau, s_\tau | \pi) Q(A)} [\ln Q(s_\tau | \pi) Q(A) - \ln P(o_\tau, s_\tau, A)] \quad (2.6)
 \end{aligned}$$

The expected free energy of a policy can be unpacked in a number of ways. Perhaps the most intuitive is in terms of risk and ambiguity:¹¹

$$G(\pi) = \underbrace{D_{KL}[Q(s_\tau, A | \pi) || P(s_\tau, A)]}_{\text{Risk}} + \underbrace{\mathbb{E}_{Q(o_\tau, s_\tau | \pi)}[-\ln P(o_\tau | s_\tau, A)]}_{\text{Ambiguity}} \quad (2.7)$$

The equivalence between the expected free energy as shown in equations 2.6 and 2.7 rests on a mean-field assumption that equates the variational posterior for states and parameters with the product of their marginal posteriors. This means that policy selection minimizes risk and ambiguity. Risk, in this setting, is simply the difference between predicted and prior beliefs about final states. In other words, policies will be deemed more likely if they bring about states that conform to prior preferences. In the optimal control literature, this part of expected free energy underwrites KL control (Todorov, 2008; van den Broek et al., 2010). In economics, it leads to risk-sensitive policies (Fleming & Sheu, 2002). Ambiguity reflects the uncertainty about future outcomes, given hidden states. Minimizing ambiguity therefore corresponds to choosing future states that generate unambiguous and informative outcomes (e.g., switching on a light in the dark).

Sometimes it is useful to express risk in terms of outcomes as opposed to hidden states—for example, when the generative model is unknown or one can only quantify preferences about outcomes (as opposed to the inferred causes of those outcomes). In these cases, the risk over hidden states can be replaced by the risk over outcomes by assuming the divergence between the predicted and true posterior is small (omitting parameters for clarity):

$$\begin{aligned}
 \underbrace{D_{KL}[Q(s_\tau | \pi) || P(s_\tau)]}_{\text{Risk (states)}} &= \underbrace{D_{KL}[Q(o_\tau | \pi) || P(o_\tau)]}_{\text{Risk (outcomes)}} \\
 &+ \underbrace{\mathbb{E}_{Q(o_\tau | \pi)}[D_{KL}[Q(s_\tau | o_\tau, \pi) || P(s_\tau | o_\tau)]]}_{\text{Expected evidence bound}} \quad (2.8)
 \end{aligned}$$

¹¹The appendix provides derivations of equation 2.7 based on the principles of optimal Bayesian design and an integral fluctuation theorem described in Friston (2019).

This divergence constitutes an expected evidence bound that also appears if we unpack expected free energy in terms of intrinsic and extrinsic value:¹²

$$\begin{aligned}
 G(\pi) &= - \underbrace{\mathbb{E}_{Q(o_\tau|\pi)}[\ln P(o_\tau)]}_{\text{Extrinsic value}} + \underbrace{\mathbb{E}_{Q(o_\tau|\pi)}[D_{KL}[Q(s_\tau, A|o_\tau, \pi)||P(s_\tau, A|o_\tau)]]}_{\text{Expected evidence bound}} \\
 &\quad - \underbrace{\mathbb{E}_{Q(o_\tau|\pi)}[D_{KL}[Q(s_\tau|o_\tau, \pi)||Q(s_\tau|\pi)]]}_{\text{Intrinsic value (states) or salience}} \\
 &\quad - \underbrace{\mathbb{E}_{Q(o_\tau, s_\tau|\pi)}[D_{KL}[Q(A|o_\tau, s_\tau, \pi)||Q(A)]]}_{\text{Intrinsic value (parameters) or novelty}} \\
 &\geq - \underbrace{\mathbb{E}_{Q(o_\tau|\pi)}[\ln P(o_\tau)]}_{\text{Expected log evidence}} - \underbrace{\mathbb{E}_{Q(o_\tau|\pi)}[D_{KL}[Q(s_\tau, A|o_\tau, \pi)||Q(s_\tau, A|\pi)]]}_{\text{Expected information gain}}
 \end{aligned}
 \tag{2.9}$$

The inequality in the final line of equation 2.9 is obtained by omitting the expected evidence bound that appears on the previous lines. As a KL-divergence, this is never negative and so ensures the final line is never greater than the expected free energy. In addition, the intrinsic value terms have been combined into the intrinsic value of both parameters and states. Extrinsic value is just the expected value of log prior preferences (i.e., log evidence), which can be associated with reward and utility in behavioral psychology and economics, respectively (Barto, Mirolli, & Baldassarre, 2013; Kauder, 1953; Schmidhuber, 2010). In this setting, extrinsic value is the complement of Bayesian risk (Berger, 2011). The intrinsic value of a policy is its epistemic value or affordance (Friston et al., 2015). This is just the expected information gain afforded by a particular policy, which can be about hidden states (i.e., salience) or model parameters (i.e., novelty). It is this term that underwrites artificial curiosity (Schmidhuber, 2006). The final inequality above shows that extrinsic value is the expected log evidence under beliefs about final outcomes, while the intrinsic value ensures that this expectation is maximally informed when outcomes are encountered. Collectively, these two terms underwrite the resolution of uncertainty about hidden states (i.e., information gain) and outcomes (i.e., expected surprisal) in relation to prior beliefs.

Intrinsic value is also known as intrinsic motivation in neurorobotics (Barto et al., 2013; Oudeyer & Kaplan, 2007; Ryan & Deci, 1985), the value of information in economics (Howard, 1966), salience in the visual neurosciences, and (rather confusingly) Bayesian surprise in the visual search

¹² Because the expected evidence bound cannot be less than zero, the expected free energy of a policy is always greater than the (negative) expected intrinsic value (i.e., log evidence) plus the intrinsic value (i.e., information gain).

literature (Itti & Baldi, 2009; Schwartenbeck, Fitzgerald, Dolan, & Friston, 2013; Sun, Gomez, & Schmidhuber, 2011). In terms of information theory, intrinsic value is mathematically equivalent to the expected mutual information between hidden states in the future and their consequences, consistent with the principles of minimum redundancy or maximum efficiency (Barlow, 1961, 1974; Linsker, 1990). Finally, from a statistical perspective, maximizing intrinsic value (i.e., salience and novelty) corresponds to optimal Bayesian design (Lindley, 1956) and machine learning derivatives, such as active learning (MacKay, 1992). On this view, active learning is driven by novelty—namely, the information gain afforded to beliefs about model parameters, given future states and their outcomes. Heuristically, this curiosity resolves uncertainty about “what would happen if I did that?” (Schmidhuber, 2010). Figure 1 illustrates the compass of expected free energy, in terms of its special cases, ranging from optimal Bayesian design through to Bayesian decision theory.

3 Sophisticated Inference

So far, we have considered generative models of policies—namely, a fixed number of ordered action sequences. These generative models can be regarded as placing priors over actions that stipulate a small number of allowable action sequences. In what follows, we consider more general models, in which the random variables are actions at each point in time, such that policies become a prior over transitions among action or control states. If we relax this prior, such that successive actions are conditionally independent, we can simplify belief updating, and implicit planning, at the expense of having to consider a potentially enormous number of policies.

The simplification afforded by assuming actions are conditionally independent follows because both actions and states become Markovian. This means we can use belief propagation (Winn & Bishop, 2005; Yedidia, Freeman, & Weiss, 2005) to update posterior beliefs about hidden states and actions, given each new observation. In other words, we no longer need to evaluate the posterior over hidden states in the past to evaluate a posterior over policies. Technically, this is because policies introduced a semi-Markovian aspect to belief updating by inducing conditional dependencies between past and future hidden states. The upshot of this is that one can use posterior beliefs from the previous time step as empirical priors for hidden states and actions at the subsequent time step. This is formally equivalent to the forward pass in the forward-backward algorithm (Ghahramani & Jordan, 1997), where the empirical prior over hidden states depends on the preceding (i.e., realized) action. Put simply, we are implementing a Bayesian filtering scheme in which observations are generated by action at each time step. Crucially, the next action is sampled from an empirical prior based on (a free energy functional of) posterior beliefs about the current hidden state.

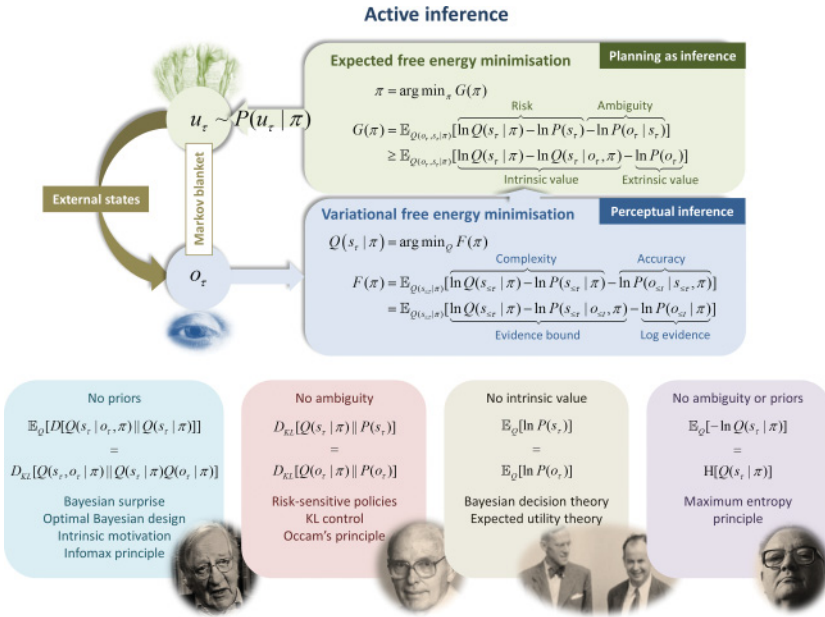


Figure 1: Active inference. This figure illustrates the various ways in which minimizing expected free energy can be unpacked. The upper panel casts action and perception as the minimization of variational and expected free energy, respectively. Crucially, active inference introduces beliefs over policies that enable a formal description of planning as inference (Attias, 2003; Botvinick & Toussaint, 2012; Kaplan & Friston, 2018). In brief, posterior beliefs about hidden states of the world, under plausible policies, are optimized by minimizing a variational (free energy) bound on log evidence. These beliefs are then used to evaluate the expected free energy of allowable policies, from which actions can be selected (Friston, FitzGerald et al., 2017). Crucially, expected free energy subsumes several special cases that predominate in psychology, machine learning, and economics. These special cases are disclosed when one removes particular sources of uncertainty from the implicit optimization problem. For example, if we ignore prior preferences, then the expected free energy reduces to information gain (Lindley, 1956; MacKay, 2003) or intrinsic motivation (Barto et al., 2013; Oudeyer & Kaplan, 2007; Ryan & Deci, 1985). This is mathematically the same as expected Bayesian surprise and mutual information that underwrites salience in visual search (Itti & Baldi, 2009; Sun et al., 2011) and the organization of our visual apparatus (Barlow, 1961, 1974; Linsker, 1990; Optican & Richmond, 1987). If we now remove risk but reinstate prior preferences, one can effectively treat hidden and observed (sensory) states as isomorphic. This leads to risk sensitive policies in economics (Fleming & Sheu, 2002; Kahneman & Tversky, 1979) or KL control in engineering (van den Broek et al., 2010). Here, minimizing risk corresponds to aligning predicted outcomes to preferred outcomes. If we then remove intrinsic value, we are left with extrinsic value or

Note that we do not need to evaluate a posterior over action, because action is realized before the next observation is generated. In other words, we can sample realized actions from an empirical prior over actions that inherits from the posterior over all previous states. This leads to a simple belief-propagation scheme for planning as inference that can be expressed as follows:

$$\begin{aligned}
 Q(s_\tau | u_{<\tau}) &= P(s_\tau | o_{<\tau}, u_{<\tau}) = \mathbb{E}_{Q(s_{\tau-1})}[P(s_\tau | s_{\tau-1}, u_{\tau-1})] \\
 Q(s_\tau) &= P(s_\tau | o_{\leq\tau}, u_{<\tau}) \propto P(o_\tau | s_\tau) Q(s_\tau | u_{<\tau}) \\
 Q(u_\tau) &= \sigma[-G(u_\tau)] \\
 G(u_\tau) &= \mathbb{E}_{P(o_{\tau+1} | s_{\tau+1}) Q(s_{\tau+1} | u_{<\tau+1})} [\\
 &\quad \underbrace{\ln Q(s_{\tau+1} | u_{<\tau+1}) - \ln P(s_{\tau+1})}_{\text{Risk}} - \underbrace{\ln P(o_{\tau+1} | s_{\tau+1})}_{\text{Ambiguity}}] \tag{3.1} \\
 &\quad \underbrace{\hspace{10em}}_{\text{Expected free energy of next action}}
 \end{aligned}$$

Here, $Q(s_\tau | u_{<\tau})$ denotes an empirical prior—from the point of view of state estimation—or a predictive posterior—from the point of view of action selection—over hidden states, given realized actions $u_{<\tau}$. Similarly, $Q(s_\tau)$ denotes the corresponding posterior, given subsequent outcomes. The first line follows immediately from the operation of marginalization, the second is an application of Bayes’s theorem, and the third is from equation 2.6. This scheme is exact because we have made no mean-field approximations of the sort required by variational message passing (Dauwels, 2007; Friston, Parr, et al., 2017; Parr et al., 2019; Winn & Bishop, 2005). Note that $Q(s_1 | u_{<1}) = P(s_1)$, with all subsequent Q distributions derived recursively

expected utility in economics (Von Neumann & Morgenstern, 1944) that underwrites reinforcement learning and behavioral psychology (Sutton & Barto, 1998). Bayesian formulations of maximizing expected utility under uncertainty are also known as Bayesian decision theory (Berger, 2011). Finally, if we just consider a completely unambiguous world with uninformative priors, expected free energy reduces to the negative entropy of posterior beliefs about the causes of data, in accord with the maximum entropy principle (Jaynes, 1957). The expressions for variational and expected free energy correspond to those described in the main text (omitting model parameters for clarity). They are arranged to illustrate the relationship between complexity and accuracy, which become risk and ambiguity, when considering the consequences of action. This means that risk-sensitive policy selection minimizes expected complexity or computational cost (Sengupta & Friston, 2018). The faces shown are, from left to right, H. Barlow, W. H. Fleming, D. Kahneman, A. Tversky, and E. T. Jaynes.

from this, meaning no variational approximation is required. However, it is worth noting a subtle difference between the Q distributions used here, and those encountered in equation 2.1). The difference is that equation 3.1 only takes account of those outcomes acquired at or before the time associated with the state. In equation 2.1), the posteriors depend on all the outcomes collected, that is, smoothing as opposed to the filtering in equation 3.1. The difference between these largely dissolves when dealing with beliefs about future states (when all relevant outcomes are earlier). Furthermore, there are no conditional dependencies on policies, which have been replaced by realized actions. However, equation 3.1 only considers the next action. The question now arises: How many future actions should we consider?

At this point, the cost of the Markovian assumption arises: if we choose a policy horizon that is too far into the future, the number of policies could be enormous. In other words, we could effectively induce a deep tree search over all possible sequences of future actions that would be computationally prohibitive. However, we can now turn to sophisticated schemes to finesse the combinatorics. This rests on the straightforward observation that if we propagate beliefs and uncertainty into the future, we only need to evaluate policies or paths that have a nontrivial likelihood of being pursued. This selective search over plausible paths is constrained at two levels. First, by propagating probability distributions, we can restrict the search over future outcomes—for any given action at any point in the future—that have a nontrivial posterior probability (e.g., greater than $1/16$). Similarly, we only need to evaluate those policies that are likely to be pursued—namely, those with an expected free energy that renders their prior probability nontrivial (e.g., greater than $1/16$).

This deep search involves evaluating all actions under all plausible outcomes so that one can perform counterfactual belief updating at each point in time (given all plausible outcomes). However, it is not necessary to evaluate outcomes per se; it is sufficient to evaluate distributions over outcomes, conditioned on plausible hidden states. This is a subtle but important aspect of finessing the combinatorics of belief propagation into the future and rests on having a generative model (that generates outcomes).

Heuristically, one can imagine searching a tree with diverging branches at successive times in the future but terminating the search down any given branch when the prior probability of an action (and the predictive posterior probability of its subsequent outcome) reaches a suitably small threshold (Keramati, Smittenaar, Dolan, & Dayan, 2016; Solway & Botvinick, 2015). To form a marginal empirical prior over the next action, one simply accumulates the average expected free energy from all the children of a given node in the tree recursively. A softmax function of this accumulated average then constitutes the empirical prior over the next action. Algorithmically, this can be expressed as follows, based on appendix 6 (Friston, FitzGerald et al. 2017), where u_τ denotes action at $\tau \geq t$ (omitting novelty terms associated with model parameters for clarity):

$$\begin{aligned}
 G(o_\tau, u_\tau) &= \mathbb{E}_{P(o_{\tau+1}|s_{\tau+1})Q(s_{\tau+1}|u_{<\tau+1})} [\\
 &\quad \underbrace{\ln Q(s_{\tau+1}|u_{<\tau+1}) - \ln P(s_{\tau+1})}_{\text{Risk}} - \underbrace{\ln P(o_{\tau+1}|s_{\tau+1})}_{\text{Ambiguity}} \\
 &\quad \underbrace{\hspace{10em}}_{\text{Expected free energy of next action}} \\
 &\quad + \underbrace{\mathbb{E}_{Q(u_{\tau+1}|o_{\tau+1})Q(o_{\tau+1}|u_{\le\tau})} [G(o_{\tau+1}, u_{\tau+1})]}_{\text{Expected free energy of subsequent actions}} \\
 Q(u_\tau|o_\tau) &= \sigma[-G(o_\tau, u_\tau)] \\
 Q(o_\tau|u_{<\tau}) &= \mathbb{E}_{Q(s_\tau|u_{<\tau})} [P(o_\tau|s_\tau)] \tag{3.2}
 \end{aligned}$$

Posterior beliefs over hidden states and empirical priors over action are then recovered from the above recursion as follows, noting that one’s most recent action (u_{t-1}) and current outcome (o_t) are realized (i.e., known) variables:

$$\begin{aligned}
 Q(s_t) &\propto P(o_t|s_t)Q(s_t|u_{<t}) \\
 Q(u_t) &= \sigma[-G(o_t, u_t)] \tag{3.3}
 \end{aligned}$$

Equation 3.3 expresses the expected free energy of each potential next action (u_τ) as the risk and ambiguity of that action plus the average expected free energy of future beliefs, under counterfactual outcomes and actions ($u_{\tau+1}$). Readers familiar with the Bellman optimality principle (Bellman, 1952) may recognize a formal similarity between equation 3.2 and the Bellman equation because both inherit from the same recursive logic. The sophisticated inference scheme deals with functionals (functions of belief distributions over states), while the Bellman equation deals directly with functions of states.

Figure 2 provides a schematic that casts this recursive formulation as a deep tree search. This search can be terminated at any depth or horizon. Later, we will rewrite this recursive scheme in terms of sufficient statistics to illustrate its simplicity. It would be possible to formulate each path through the tree of actions as an alternative policy and simply sum the expected free energy, based on current posterior beliefs, along each of those paths. This is the approach that has traditionally been pursued in active inference (Friston et al., 2016; Friston, FitzGerald et al., 2017), and accounts for the consequences of action on belief updating. The advance offered by the sophisticated formulation is that it also accounts for the consequences of anticipated belief updates for future actions. In other words, an unsophisticated creature may entertain the belief that *if I did that, I would find out about this*. A sophisticated creature additionally believes that *if I found that out, I would then do this*. An intuitive example would be in deciding whether to check the news, look at the weather forecast, read a novel, or go for a walk. The first two options might offer similar information gain and would

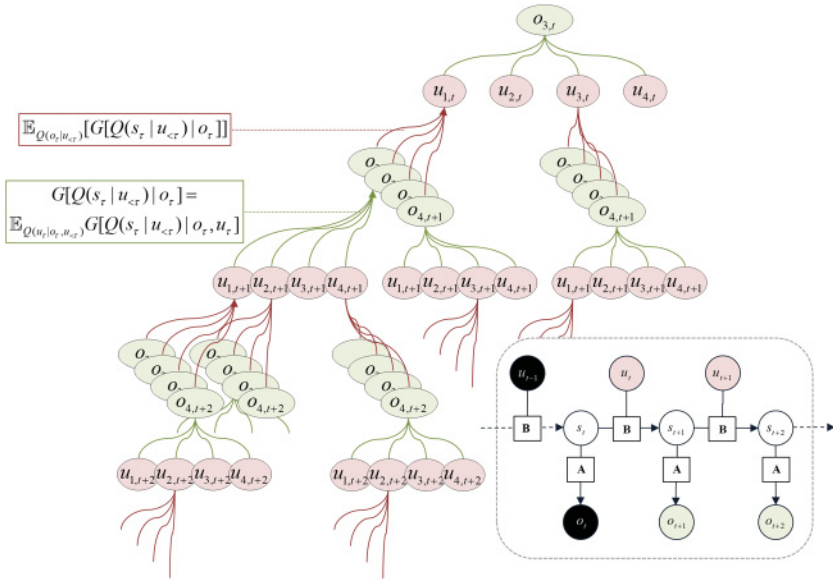


Figure 2: Deep policies searches. This schematic summarizes the accumulation of expected free energy over paths or trajectories into the future. This can be construed as a deep tree search, where the tree branches over allowable actions at each point in time and the likely outcomes consequent on each action. The arrows between actions and outcomes have been drawn in the reverse direction (directed from the future) to depict the averaging of expected free energy over actions (green arrows) and subsequent averaging over the outcomes entailed by the preceding action (pink arrows). This dual averaging over actions (given outcomes) and outcomes (given actions) is depicted by the equations in the upper panel. Here, the green nodes of this tree correspond to outcomes, with one (realized) outcome at the current time (at the top). The pink nodes denote actions—here, just four. Note that the search terminates whenever an action is deemed unlikely or an outcome is implausible. The panel on the lower right represents the conditional dependencies in the generative model as a probabilistic graphical model. The parameters of this model are shown on squares, and the variables are shown on circles. The arrows denote conditional dependencies. Filled circles are realized variables at the current time—namely, the preceding action and the subsequent outcome. Note that the expected free energy is shown here as a functional of beliefs about states, where these beliefs are updated based on actions and outcomes. In the main text, we drop the explicit dependence on Q and express the expected free energy directly as a function of outcomes and actions.

appeal to an unsophisticated agent. Without knowing the weather, the latter two will be hard to disambiguate given preferences for walking in the sun or reading indoors if it were raining. A more sophisticated agent will find the weather forecast more salient than the news: knowing the weather will determine whether the next action will be to go for a walk or stay in and read, given that the preferred option is more likely to be chosen once the weather is known.

This sort of approach to evaluating a tree of possible policies, using a recursive form for the expected free energy, has been suggested by others (Çatal, Verbelen, Nauta, Boom, & Dhoedt, 2020; Çatal, Wauthier, et al., 2020), who have applied this in the context of robot vision and navigation. The distinction between this and the formulation presented here is the sophisticated aspect: here, each additional step into the future evaluates the expected free energy in terms of the beliefs anticipated at that time point, as opposed to beliefs held (at the present) about that time point. Despite this difference, the similarities in these approaches speak to the feasibility of scaling sophisticated inference to high-dimensional

Having established the formal basis of sophisticated planning, in terms of belief propagation, we now turn to some illustrative examples to show how it works in practice.

4 Simulations

In this section, we provide some simulations to compare sophisticated and unsophisticated schemes on the three-arm bandit task described in section 1. Here, we frame this paradigm in terms of a rat foraging in a three-arm T-maze, where the right and left upper arms are baited with rewards and punishments, and the bottom arm contains an instructional cue indicating whether the bait is likely to be on the right or left. In these examples, cue validity was 95%. The details of this setup have been described elsewhere (Friston et al., 2016; Friston, FitzGerald et al., 2017). In brief, the generative model comprises a likelihood mapping between hidden states and outcomes and probability transitions among states. Here, there are two outcome modalities. The first reports the experience of the rat in terms of its location (with distinct outcomes for the instructional cue location – *right* versus *left*). The second modality registered rewarding outcomes, with three levels (*none*, *reward*, and *punishment*—for example, foot shock). There were two hidden factors: the rat’s location (with four possibilities) and the latent context (i.e., whether the rewarding arm was on the right or the left). With these hidden states and outcomes, we specify the generative model in terms of:

- The sensory mapping \mathbf{A} , which maps from the two hidden state factors (location and context) to each of the two sensory modalities (location and reward).

- The transition matrices \mathbf{B} , which govern how states at one time point map onto the next, given a particular action (u_t). The transitions among locations are action dependent, with four actions (moving to one of the four locations), while the context did not change during any particular trial (i.e., there were no context transitions within trials).
- The cost vectors \mathbf{C} for each hidden state factor, which also specify the agent's preferences for each outcome modality. The latter allows for an alternative formulation that we discuss below.
- The priors over initial states, \mathbf{D} .

In the following simulations, the rat experienced 32 trials, each comprising two moves with three outcomes, including an initial outcome that located the rat at the start (i.e., center) location. The rat encountered the first trial with ambiguous prior beliefs about the context, that is, the reward was equally likely to be right or left.

Given this parameterization of the generative model, the expected free energy of an action, given outcomes, equation 4.1 can be expressed in terms of sufficient statistics of posterior beliefs and model parameters as follows:¹³

$$\mathbf{G}(u_\tau, o_\tau) = \underbrace{\mathbf{s}_{\tau+1}^u \cdot [\ln \mathbf{s}_{\tau+1}^u + \mathbf{C} + \mathbf{H}]}_{\text{Expected free energy of next action}} + \underbrace{\mathbf{u}_{\tau+1}^o \cdot \mathbf{G}(u_{\tau+1}, o_{\tau+1}) \mathbf{o}_{\tau+1}^u}_{\text{and subsequent actions}}$$

$$\mathbf{s}_\tau \propto (\mathbf{A} \cdot o_\tau) \odot \mathbf{s}_\tau^u$$

$$\mathbf{s}_\tau^u = \mathbf{B}(u_{\tau-1}) \mathbf{s}_{\tau-1}$$

$$\mathbf{o}_\tau^u = \mathbf{A} \mathbf{s}_\tau^u$$

$$\mathbf{u}_\tau^o = \sigma[\mathbf{G}(u_\tau, o_\tau)] \quad (4.1)$$

Here, \odot denotes a Hadamard (i.e., element-wise) product, and the dot notation means $\mathbf{A} \cdot o_\tau \equiv \mathbf{A}^T o_\tau$. \mathbf{H} is the conditional entropy of the likelihood distribution. The sufficient statistics are the parameters of the categorical distributions in equation 3.2, where model parameters are usually hyperparameterized in terms of the concentration parameters of Dirichlet distributions (denoted by capital and lowercase bold variables, respectively):

$$Q(\mathbf{s}_\tau) = \text{Cat}(\mathbf{s}_\tau)$$

$$Q(\mathbf{s}_{\tau+1} | u_{\leq \tau}) = \text{Cat}(\mathbf{s}_{\tau+1}^u)$$

$$Q(u_\tau | o_\tau) = \text{Cat}(\mathbf{u}_\tau^o)$$

¹³We have suppressed any tensor notation here by assuming there is only one outcome modality and one hidden factor. In practice, this assumption can be guaranteed by working with the Kronecker tensor product of hidden factors. This ensures exact Bayesian inference, because conditional dependencies among hidden factors are evaluated.

$$\begin{aligned}
 Q(o_\tau | u_{<\tau}) &= \text{Cat}(\mathbf{o}_\tau^u) \\
 P(o_\tau | s_\tau) &= \text{Cat}(\mathbf{A}) \\
 P(s_{\tau+1} | s_\tau, u_\tau) &= \text{Cat}(\mathbf{B}(u_\tau)) \\
 P(s_1) &= \text{Cat}(\mathbf{D}) \\
 \mathbf{C} &= -\ln P(s_\tau) \\
 \mathbf{H} &= -\text{diag}(\mathbf{A} \cdot \ln \mathbf{A}) \\
 P(\mathbf{A}) &= \text{Dir}(a) \\
 P(\mathbf{B}) &= \text{Dir}(b) \\
 P(\mathbf{D}) &= \text{Dir}(d)
 \end{aligned} \tag{4.2}$$

The equivalent scheme, when specifying preferences in terms of outcomes $\mathbf{C} = \ln P(o_\tau)$, is

$$\mathbf{G}(u_\tau, o_\tau) = \underbrace{\mathbf{o}_{\tau+1}^u \cdot [\ln \mathbf{o}_{\tau+1}^u + \mathbf{C}] + \mathbf{s}_{\tau+1}^u \cdot \mathbf{H}}_{\text{Expected free energy of next action}} + \underbrace{\mathbf{u}_{\tau+1}^o \cdot \mathbf{G}(u_{\tau+1}, o_{\tau+1})}_{\text{and subsequent actions}} \mathbf{o}_{\tau+1}^u \tag{4.3}$$

As noted, it is usually more convenient to search over distributions over outcomes that are generated by (plausible) hidden states as opposed to (plausible) outcomes per se. This approach produces a slightly simpler form for expected free energy:

$$\mathbf{G}(u_\tau, o_\tau) = \mathbf{s}_{\tau+1}^u \cdot \underbrace{[\ln \mathbf{s}_{\tau+1}^u + \mathbf{C} + \mathbf{H}]}_{\text{Next action}} + \underbrace{\mathbf{G}(u_{\tau+1}, \mathbf{A}\mathbf{s}_{\tau+1}^u) \cdot \mathbf{u}_{\tau+1}^o}_{\text{and subsequent actions}} \tag{4.4}$$

Finally, as intimated above, the recursive estimation of expected free energy from subsequent states can be terminated when the probability of an action or outcome can be plausibly discounted. In the simulations here, searches over paths were terminated when the predictive probability fell below 1/16. This choice of threshold is a little arbitrary and could itself be optimized either in relation to the accumulated free energy for a synthetic agent or in fitting empirical behavior. However, the 1/16 value offers a useful balance as it enables elimination of policies that are highly unlikely, improving efficiency of planning while also being relatively conservative. It corresponds to a probability of about 0.06, close to the ubiquitous 0.05 used to reject null hypotheses in frequentist statistics. We make no claim as to 1/16 being the optimal threshold in the context of all tasks—or even in those shown here. However, this is something that could be optimized in relation to a specific

task by finding the threshold that minimizes the free energy accumulated over time.

While crude, this works under the assumption that if one policy is 16 times less likely than alternatives given how far it has been evaluated, it is unlikely to be redeemed by evaluating it further. As such, there are savings to be had in not doing so. If there were no constraints on computational resources (temporally or thermodynamically), the pruning threshold could be set to be zero, ensuring an exhaustive evaluation of all possible policies. The principles that underwrite sophisticated inference do not depend on this specific implementational detail, and alternative methods could be used.

Other approaches to searching through policy trees include schemes like Thompson sampling (Ortega & Braun, 2010; Osband, Van Roy, Russo, & Wen, 2019; Thompson, 1933), which sample from the posterior probability for states and select policies that maximize preferences given this sample. Like the threshold we have selected, this simplifies the search through alternative policies by using samples in place of evaluating the full posterior probabilities. With enough exposure to a task, Thompson sampling ensures that the full space of plausible policies is attempted, possibly finding “optimal” policies that are discounted by early pruning under our approach. In our setting, Thompson sampling would not be appropriate because our focus is on inference (selecting the best policy within a trial) as opposed to learning a policy over many exposures to a trial. Having said this, it is worth highlighting that action selection using the sophisticated inference scheme involves sampling from the posterior distribution over actions—subject to some temperature parameter. While this parameter is typically very large so that the maximum a posteriori action is chosen, this could be relaxed to ensure the occasional selection of unlikely actions, in the spirit of Thompson sampling.

The simulations were chosen to illustrate the fidelity of beliefs about action (i.e., what to do next) with and without a sophisticated update scheme (see equations 3.1 and 3.2). We anticipated that sophisticated schemes would outperform unsophisticated schemes, in the sense that they would learn any contingencies more efficiently, via more confident action selection. This learning was elicited by baiting the left arm consistently, after a couple of trials, so that priors about the initial (latent context) state could be accumulated, in the form of posterior (Dirichlet) concentration parameters (\mathbf{d}). In these generative models, learning is straightforward and involves the accumulation of posterior concentration parameters (Friston et al., 2016). For example, to learn the likelihood mapping and initial hidden states, we have¹⁴

¹⁴Note that in order to accumulate beliefs about the context from trial to trial, it is necessary to carry over posterior beliefs about context from one trial as prior beliefs for

$$\begin{aligned} \mathbf{A} &= \mathbf{a} \odot \mathbf{a}_0^{\odot -1}, & \mathbf{a}_{0ij} &= \sum_i \mathbf{a}_{ij}, & \mathbf{a} &= \sum_{\tau} a + o_{\tau} \otimes \mathbf{s}_{\tau} \\ \mathbf{D} &= \mathbf{d} \odot \mathbf{d}_0^{\odot -1}, & \mathbf{d}_{0i} &= \sum_i \mathbf{d}_i, & \mathbf{d} &= d + \mathbf{s}_1 \end{aligned} \quad (4.5)$$

In these sorts of simulation, the agent succumbs to the epistemic affordance of the instructional cues until it learns that the reward is always on the left-hand side—at which point, the expected utility (or extrinsic value) of going directly to the baited arm exceeds the epistemic affordance (or intrinsic value) of soliciting the instructional cue. At this point, there is a switch from explorative to exploitative behavior—the behavioral measure we used to compare sophisticated and unsophisticated schemes.

4.1 Exploration and Exploitation in a T-Maze. Figure 3 shows the results of three simulations. In these simulations, the rat performed 32 trials where each trial had two moves, starting from the central location. The prior preferences for reward and punishment outcomes were specified with the prior costs (C) of -2 and 2 , respectively.¹⁵ In these and subsequent simulations, actions were selected as the most likely (maximum a posteriori) action. Therefore, all subsequent simulations are deterministic realizations of (Bayes's) optimal behavior based on expected free energy. The simulations start with a sophisticated agent with a planning horizon of two (this corresponds to the depth of action sequences considered into the future). In other words, it accumulates the expected free energy for all plausible paths, until the end of each trial. This enables a confident and definitive epistemic policy selection that gives way to exploitation, when the rat realizes the reward is always located in the left arm.

If we compare this performance with that of an unsophisticated rat, which looks just one move ahead, we see a similar behavior. However, there are two differences. First, the rat is less confident about its behavior because it does not evaluate the consequences of its actions in terms of belief updating. Although it finds the instructional cue more attractive, in virtue of its epistemic affordance, it is still partially compelled to remain at the central location, which ensures that it will avoid aversive outcomes. Because the

the next (in the form of Dirichlet concentration parameters). For consistency with earlier formulations of this paradigm, we carry over the beliefs about the initial state on the previous trial that are evaluated using a conventional backwards pass—namely, the normalized likelihood of any given initial state, given subsequent observations—and probability transitions based on realized action.

¹⁵ Because costs are specified in terms of self-information or surprisal, they have meaningful and quantitative units. For example, a differential cost of three natural units corresponds to a log odds ratio 1:20 and reflects a strong preference for one state or outcome over another. This is the same interpretation of Bayes factors in statistics (Kass & Raftery, 1995) Here, the difference between reward and punishment was four natural units.

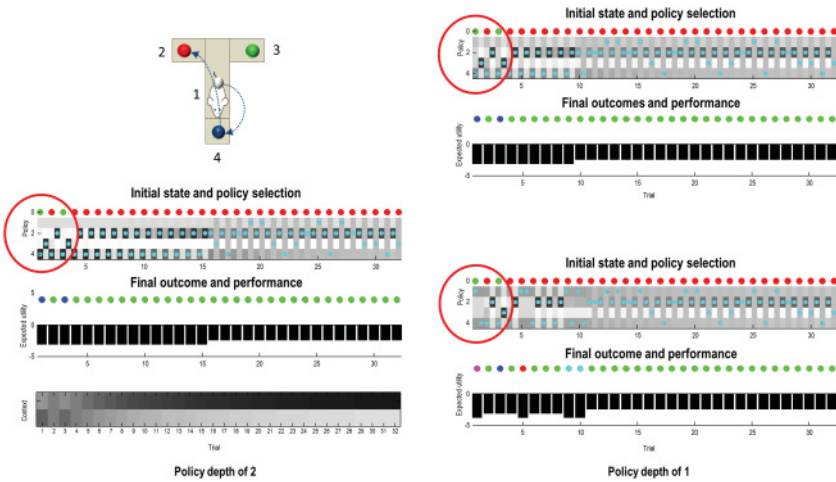


Figure 3: Epistemic foraging in a T-maze: This figure shows the results of simulations based on the T-maze paradigm described in the main text. The left panel shows the results of simulating 32 trials, where the rat started at the central location. Each trial comprises two moves. The insert on the upper left illustrates foraging for information by interrogating the instructional cue in the lower arm and then securing the reward in the left arm. The results in each of the three panels have the same format. The upper row illustrates the predictive distribution over actions (moves to the central location, the left, the right, and lower arm, respectively). The darker the color, the more likely the action. The cyan dots are the actions that were sampled and executed at each epoch, within each trial. The colored dots above indicate the hidden context—namely, whether the left or right arm was baited. The middle panel shows the resulting performance in terms of the expected utility or negative Bayesian risk. The colored circles show the final outcome (blue location 3—right arm—and green location 2—left arm). The lowest panel (on the left) shows the posterior beliefs about the hidden context (right versus left) based on Dirichlet concentration parameters, accumulated over trials. The left panel of results shows confident epistemic behavior with a planning horizon of two. As is typical in these kinds of simulations, the agent starts off by foraging for information and responding to the epistemic affordance of the instructional cue in the lower arm. However, because the reward is always encountered in the left arm (after the first couple of trials), the rat loses interest in the instructional cue as it becomes more confident about where the reward is located. This experience-dependent loss of epistemic affordance leads to a switch from exploratory to exploitative behavior—here, at trial 16. A similar kind of behavior is shown in the upper right panels; however, here, the planning horizon was reduced to one. In other words, the rat considered only the expected free energy of one move ahead. The key difference here is a less confident (i.e., precise) belief distribution over early actions (highlighted by the red circles). Although the lower arm has the greatest posterior probability, there is a nontrivial probability that the rat thinks it should stay where it is. This

unsophisticated agent underestimates the epistemic affordance of the instructional cue, it paradoxically performs better in terms of suspending its information foraging earlier and switching to exploitative behavior a few trials before the sophisticated agent (but see below).

For completeness, we show the results of an unsophisticated agent, whose behavior is predicated on Bayesian risk, that is, with no epistemic value in play. As might be anticipated, this agent exposes itself to Bayesian risk, forgoing a visit to the right or left arm, in a way that is precluded by agents who minimize expected free energy. Here, the starting and instructional cue locations are equally attractive. When the rat is lucky enough to select the lower arm, it knows what to do; however, it has no sense that this is the right kind of behavior. After a sufficient number of trials, it realizes that the reward is always on the left-hand side and starts to respond in an exploitative fashion, albeit with relatively low confidence. These results highlight the distinction between sophisticated and unsophisticated agents who predicate their policy selection on expected free energy and between unsophisticated agents using expected free energy with and without epistemic affordance.

In the simulations, the sophisticated agent persevered with its epistemic behavior for longer than the unsophisticated agent. At first glance, this may seem to be a paradoxical result if we were measuring performance in terms of Bayesian risk. However, this is not the case as illustrated in Figure 4F. Here, we repeated the simulations above but with one small change: we made the epistemic cue mildly aversive by giving it a cost of one. This has no effect on the sophisticated agent other than slightly abbreviating the exploratory phase of activity. However, the unsophisticated agent has, understandably, been caught in a bind. The starting location is now marginally more preferable than the instructional cue—and it has no reason to leave the center of the maze. While this ensures aversive outcomes are avoided, it also precludes epistemic foraging and subsequent exploitation. Heuristically, only the sophisticated agent can see past the short-term pain for the long-term gain. We will pursue this theme in the final simulations, where the agent's planning horizon becomes nontrivial.

mild ambiguity about what should be done means that exploratory behavior yields to exploitative behavior slightly earlier, at trial 10. Finally, the lower right panels show the results when expected free energy is replaced by Bayesian risk. In other words, any epistemic affordance of the instructional cue is precluded. This renders the posterior probability of staying or moving to the lower arm the same. When, by chance, the instructional cue is encountered, exploitative behavior follows; however, there are times when the rat simply stays at the central location and learns nothing about the prevailing context. Note that in this example, there are costly trials in which the rat fails to visit either baited arm.

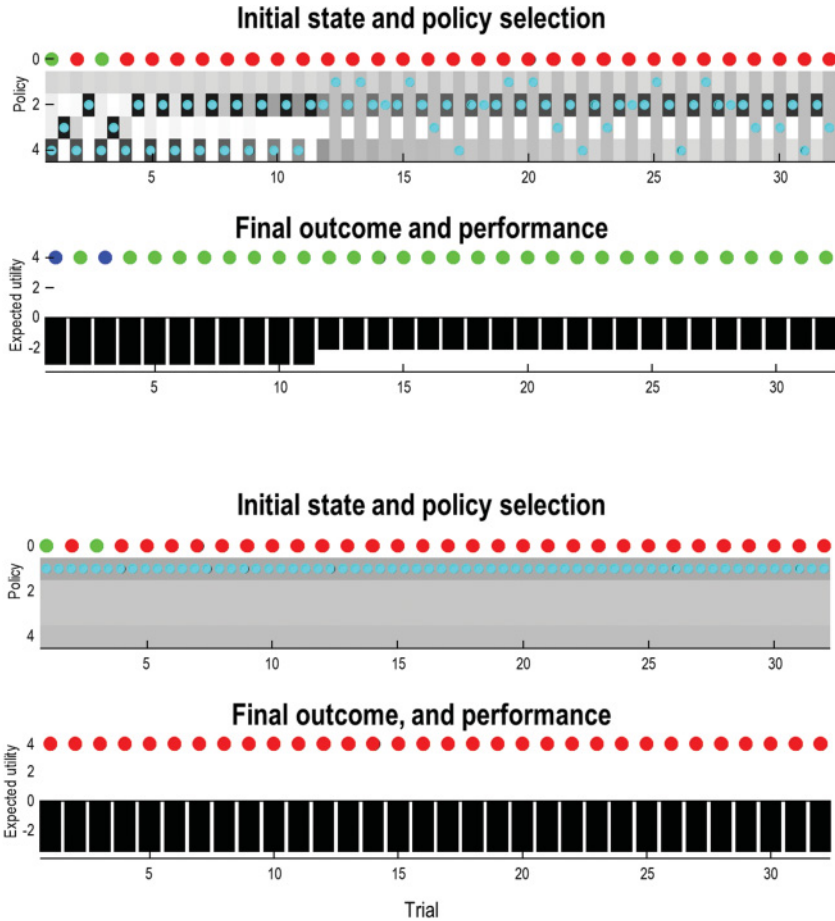


Figure 4: This reproduces the results of Figure 3 with a deep policy search (of horizon or depth 2). However, here, we have made the lower arm slightly averse. This is no problem for the sophisticated agent who sees through the short-term cost to visit the instructional cue as usual. Because this location is mildly aversive, the switch to exploitative behavior is now slightly earlier (at trial 12). Contrast this behavior with an unsophisticated agent that does not look beyond its next move. The resulting behavior is shown in the lower panels. Unsurprisingly, the agent just stays at the starting position and learns nothing about its environment—and safely avoids all adverse outcomes at the expense of forgoing any rewards.

4.2 Deep Planning and Navigation. The simulations show that a sophisticated belief-updating scheme enables more confident and nuanced policy selection, which translates into more efficient exploitative behavior. To illustrate how this scheme scales up to deeper policy searches, we revisit a problem that has been previously addressed using a bespoke prior, based on the graph Laplacian (Kaplan & Friston, 2018). This problem was previously framed in terms of navigation to a target location in a maze. Here, we forgo any special priors to see if the sophisticated scheme could handle deep tree searches that underwrite paradoxical behaviors, like moving away from a target to secure it later (see the mountain car problem). Crucially, in this instance, there was no ambiguity about the hidden states. However, there was ambiguity or uncertainty about the likelihood mapping that determines whether a particular location should be occupied. In other words, this example uses a more conventional foraging setup in which the rat has to learn about the structure of the maze while simultaneously pursuing its prior preferences to reach a target location. Here, exploratory behavior is driven by the intrinsic value or information gain afforded to beliefs about parameters of the likelihood model (as opposed to hidden states). Colloquially, one can think of this as epistemic affordance that is underwritten by novelty as opposed to salience (Barto et al., 2013; Parr & Friston, 2019a; Schwartenbeck et al., 2019). Having said this, we anticipated that exactly the same kind of behavior would arise and that the sophisticated scheme would be able to plan to learn and then exploit what it has learned.

In this paradigm, a rat has to navigate over the 8×8 grid maze, where each location may or may not deliver a mildly aversive stimulus (e.g., a foot shock). Navigation is motivated by prior preferences to occupy a target location—here, the center. In the simulations below, the rat starts at the entrance to the maze and has a prior preference for safe outcomes (cost of -1) and against aversive outcomes (cost of $+1$). Prior preferences for location depend on the distance from the current position to the target location. The generative model for this setup is simple: there was one hidden factor with 64 states corresponding to all possible locations. These hidden states generate safe or aversive (somatosensory) outcomes, depending on the location. In addition, (exteroceptive) cues are generated that directly report grid location. The five allowable actions comprise one step in any direction or staying put.

Figure 5 shows the results of typical simulations when increasing the planning horizon from 1 through to 4. The key point here is that there is a critical horizon, which enables our subject to elude local minima of expected free energy as it pursues its goal. In these simulations, our subject was equipped with full knowledge of the aversive locations and simply planned a route to its target location. However, relatively unsophisticated agents get stuck on the other side of aversive barriers that are closest to the target location. In other words, they remain in locations in which the

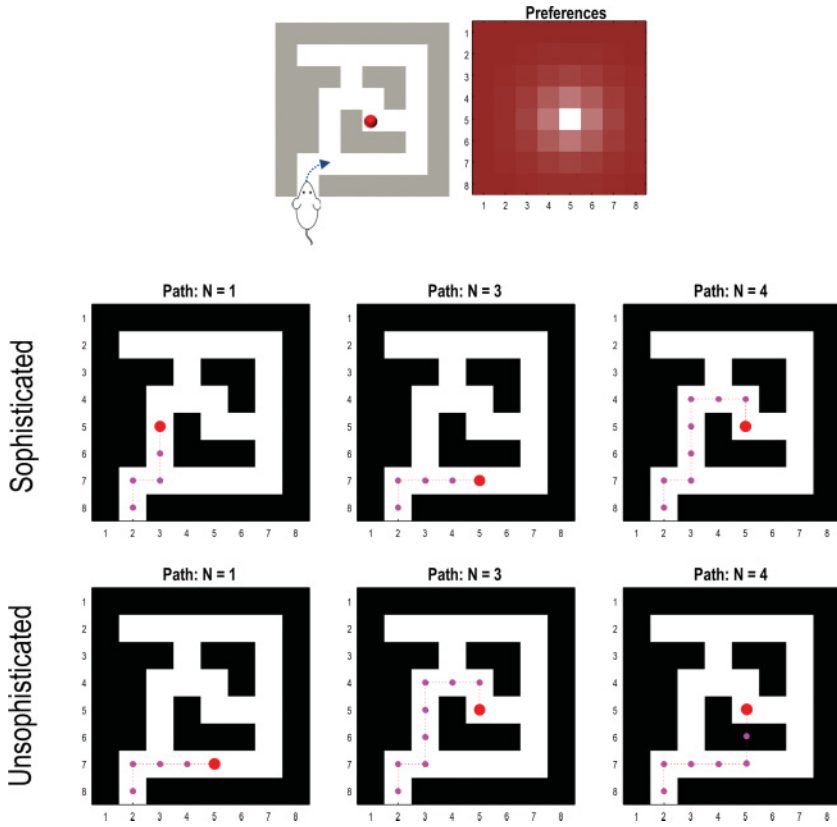


Figure 5: Navigation as inference: This figure reports the result of a simulated maze navigation. The upper panels illustrate the form of this maze, which comprises an 8×8 grid. Each location may or may not deliver a mildly aversive outcome (e.g., a foot shock). At the same time, the rat's prior preference is to be near the center of the maze. These prior preferences are shown in image format in the top right panel, where the log prior preference is illustrated in pink, with white being the most preferred location. The bottom three panels record the trajectory or path taken by a rat from the starting location on the lower left. The three panels show the (deterministic) solutions for a planning horizon of 1, 3, and 4. With horizons of fewer than four, the rat gets stuck on the other side of an aversive barrier that is closest to the central (i.e., target) location. This is because any move away from this location (with a small excursion) has a smaller expected free energy than staying put. However, if the policy search is sufficiently deep (i.e., a planning horizon greater than 3), the rat can effectively imagine what would happen if it pressed deeper into the future, enabling long-term gains to supervene over short-term losses. The result is that the rat infers and pursues the shortest path to the target location, even though it occasionally moves away from the center. The bottom three panels illustrate the behavior of

expected free energy of leaving is always greater than staying put (Cohen, McClure, & Yu, 2007). This can happen when the planning horizon is insufficient to enable the rat to contemplate distal (and potentially preferable) outcomes (as seen in the lower left and middle panels of Figure 5). However, with a planning horizon of 4 (or more), these local minima are vitiated, and the rat easily plans—and executes—the shortest path to the target. In these simulations, the total number of moves was eight, which is sufficient to reach the target via the shortest path. This sort of behavior is reminiscent of the prospective planning required to solve things like the mountain car problem. In other words, the path of least expected free energy can often involve excursions through state (and belief) space that point away from the ultimate goal.

To aid with intuition as to the evaluation of alternative policies, we explicitly evaluated some of the policies that could be chosen with a planning horizon of two. Assuming the maze layout is known, there is little uncertainty to resolve, and preferences (i.e., costs) will be the primary determinant of behavior. Starting from the maze entrance (2,8), the options are shown in Table 2.

Here, we can see that when we consider only the first step, there is a cost of +2.6 associated with choosing up and a cost of +6.0 for choosing left. Remembering that cost is formulated as a log probability; this means up is about 30 times more likely than left and suggests we do not need to evaluate policies starting with a left (which falls below the 1/16 threshold) any further. Inspection of the options for the second step of these policies and comparison with those for the policies starting with up suggests the cost incurred at the first step cannot be compensated for at the second.

For all policies surviving the 1/16 threshold, we then have to consider the next step. For the example in Table 2, we could do this simply by taking the total cost for the second step for each action and, using a softmax operator as in equation 4.1, compute the relative probability of each action and the cost incurred on averaging under these probabilities. Adding this to the

an unsophisticated agent. This is as described in Kaplan and Friston (2018) but with constant preferences as in the upper panels and variable policy depths. In the example, the planning horizon of three is sufficient for the rat to find the shortest path. However, this depends on the rat's choosing the left path at the first junction—which is not guaranteed, as four moves along the left or the right path lead to squares that are equally preferred. Consistent with this, for the policy depth of 4, the right path is chosen. After the first four moves, the rat decides to cross the aversive square to reach the target location. This four-step policy allows the rat to entertain the benefits of spending multiple steps in the target location, at the cost of a single foot shock. In these simulations, the rat knew the locations of the aversive outcomes and was motivated by minimizing Bayesian risk.

Table 2: Example Policy Evaluation.

Action	Cost (nats)		Step 2		
	Square Color	Target Proximity	Action	Cost (nats)	
				Square Color	Target Proximity
Stay at (2,8)	-1	+4.2	Up to (2,7)	-1	+3.6
			Down to (2,8)	-1	+4.2
			Left to (1,8)	+1	+5.0
			Right to (3,8)	+1	+3.6
			Stay at (2,8)	-1	+4.2
Up to (2,7)	-1	+3.6	Up to (2,6)	+1	+2.2
			Down to (2,8)	-1	+4.2
			Left to (1,7)	+1	+4.4
			Right to (3,7)	-1	+2.8
			Stay at (2,7)	-1	+3.6
Down to (2,8)	-1	+4.2	Up to (2,7)	-1	+3.6
			Down to (2,8)	-1	+4.2
			Left to (1,8)	+1	+5.0
			Right to (3,8)	+1	+3.6
			Stay at (2,8)	-1	+4.2
Left to (1,8)	+1	+5.0	Up to (1,7)	+1	+4.4
			Down to (1,8)	+1	+5.0
			Left to (1,8)	+1	+5.0
			Right to (2,8)	-1	+4.2
			Stay at (1,8)	+1	+5.0
Right to (3,8)	+1	+2.8	Up to (3,7)	-1	+2.8
			Down to (3,8)	+1	+3.6
			Left to (2,8)	-1	+4.2
			Right to (4,8)	-1	+4.2
			Stay at (3,8)	+1	+2.8

cost from the first step and repeating for all policies not eliminated by the 1/16 threshold, we arrive at the (log) probability distribution over the first action—here, favoring up.

We have characterized the degree of sophistication in terms of planning as inference. In this setting, there was no ambiguity about outcomes that would license an explanation in terms of epistemic affordance or salience of the sort that motivated behavior in the T-maze examples of section 4.1. However, we can reintroduce epistemics by introducing uncertainty about the locations that deliver aversive outcomes. Exploration now becomes driven by curiosity about the parameters of the likelihood mapping (see equation 2.9). One can illustrate the minimization of expected free energy in terms of curiosity and novelty (Barto et al., 2013; Schmidhuber, 2006) by simulating a rat that has never been exposed to the maze previously. This was implemented by setting the prior (Dirichlet) parameters of the likelihood mapping between hidden states and somatosensory outcomes to a

small value (i.e., $1/64$). In terms of sufficient statistics, the expected free energy is now supplemented with a novelty term based on posterior expectations about the likelihood mapping (Friston, Lin, et al., 2017):

$$\mathbf{G}(u_{\tau-1}, o_{\tau-1}) = \underbrace{\mathbf{s}_{\tau}^u \cdot [\ln \mathbf{s}_{\tau}^u + \mathbf{C} + \mathbf{H}]}_{\text{Next action}} - \underbrace{\mathbf{o}_{\tau}^u \cdot \mathbf{W} \mathbf{s}_{\tau}^u}_{\text{Novelty}} + \underbrace{\mathbf{u}_{\tau}^o \cdot \mathbf{G}(u_{\tau}, o_{\tau}) \mathbf{o}_{\tau}^u}_{\text{Subsequent actions}}$$

$$\mathbf{W} = \frac{1}{2}(\mathbf{a}^{\odot-1} - \mathbf{a}_0^{\odot-1}) \quad (4.6)$$

In addition, we removed preferences for a particular location in order to study purely exploratory behavior. The results of the ensuing simulation are shown in Figure 6. In this example, the rat was allowed to make 64 consecutive moves while updating the Dirichlet parameters after every move. The top panels 6 show the resulting trajectory. The key point to observe here is that nearly every location has been explored. This rests on a trajectory in which previously visited locations lose their novelty or epistemic affordance, thereby promoting policies that take the rat into uncharted territory. This kind of exploratory behavior disappears if we replace expected free energy with Bayesian risk. In this setting, after the first move, the rat returns to its original location and just sits there for 64 trials (see the bottom panels of Figure 6).

Finally, to simulate curiosity under a task set, we reinstated prior preferences about location. In this simulation, the rat has to resolve the dual imperative to satisfy its curiosity, while at the same time realizing preferences for being at the center of the maze. In other words, it has to contextualize its goal-seeking behavior in relation to what it knows about how to realize those goals. Figure 7 shows the results of a simulation in which the rat was given five exposures to the maze, each comprising eight moves with a planning horizon of four. Within four exposures, it has learned what it needs to learn—about the aversive locations—to plan the shortest path to its target location and execute that path successfully (dotted black line in the left panel of Figure 7). In contrast to Figure 6, the exploration is now limited to preferred locations with precise likelihood mappings that are sufficient to encompass the shortest path (compare the left panels of Figures 6 and 7).

This completes our numerical analyses, in which we have looked at deep policy searches predicated on expected free energy, where expected free energy supplements Bayesian risk with epistemic affordance in terms of either salience (resolving uncertainty about hidden states) or novelty (resolving uncertainty about hidden model parameters).

5 Conclusion

This letter has described a recursive formulation of expected free energy that effectively instigates a deep tree search for planning as inference. The ensuing planning is sophisticated, in the sense that it entails beliefs

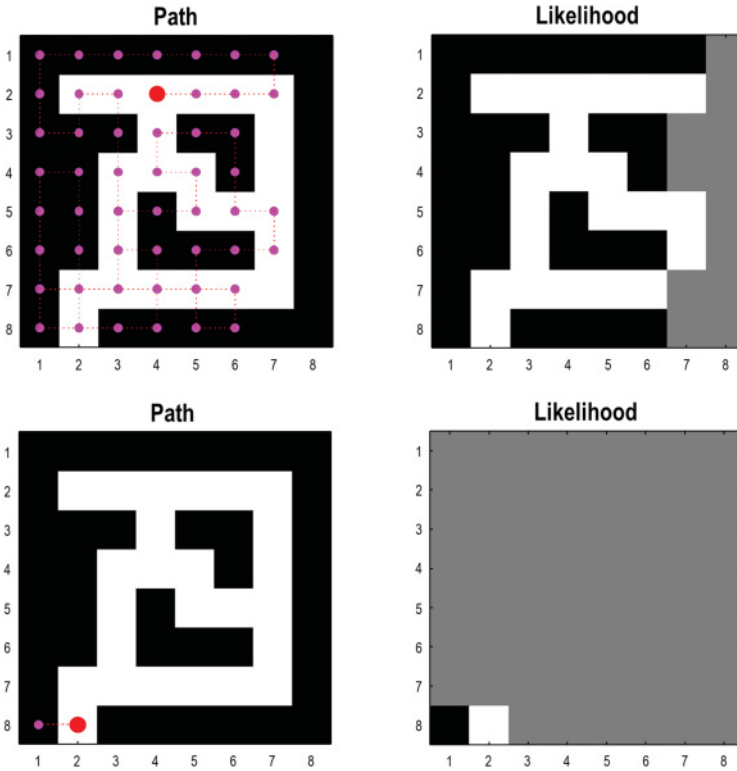


Figure 6: Exploration and novelty: This figure reports the results of a simulation in the same maze as in Figure 5. However, here we removed prior knowledge about which locations should be avoided and prior preferences for being near the center. This means that the only incentives for movement are purely epistemic in nature: curiosity, or the novelty of finding out “what would happen if I did that.” This produces a trajectory of moves that explore the locations, building up a picture of where aversive (a foot shock) stimuli are elicited and where they are not. The key aspect of this trajectory is that it avoids revisiting previously explored locations, to provide a nearly optimal coverage of the exploration space. The number of moves was 64 (with an updating of the posterior beliefs about likelihood parameters after each move). This means that in principle, the rat could have visited every location. Indeed, nearly every location has been visited, as shown on the upper right, in terms of the final likelihood of receiving an aversive stimulus at each location. The bottom panels show the same results, but after replacing expected free energy (that includes the novelty term) with Bayesian risk (that does not). Unsurprisingly, the Bayesian risk agent has no imperative to move, because it has no preferences about its location and, after the first move, realizes it is in a safe location. In other words, after the first move, it returns to the starting location and remains there for the remainder of available trials. As such, it learns nothing about the mapping between location and sensory outcomes.

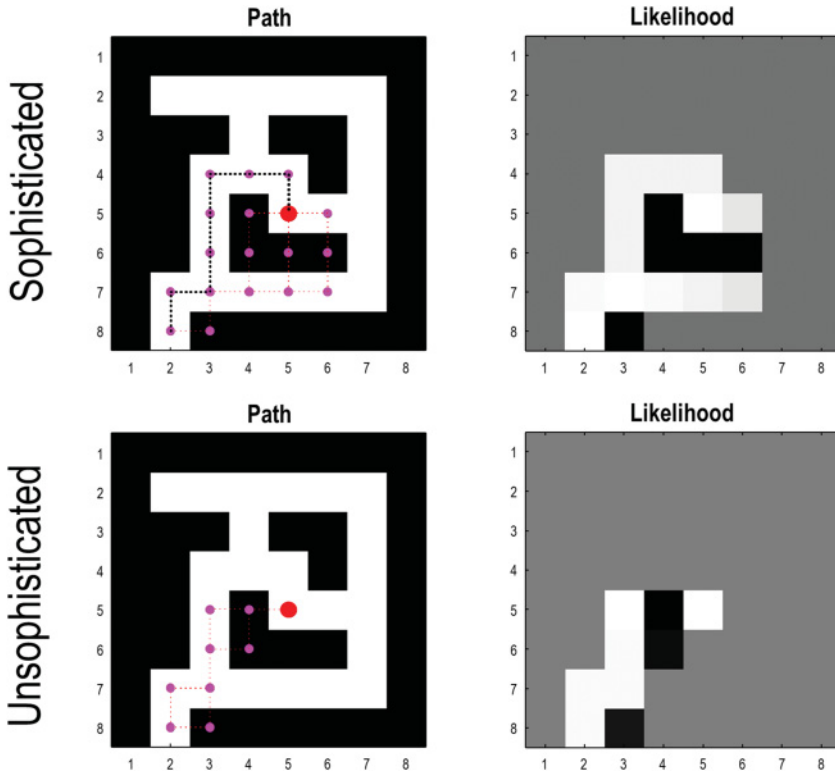


Figure 7: Exploration under a task set: This figure reproduces the same paradigm as in Figure 6 but reinstating prior preferences about being near the center of the maze (i.e., a task set). In this instance, the imperatives for action include both curiosity and pragmatic drives to realize prior preferences. The upper left panel shows a sequence of trajectories over five trials, where the rat was replaced at the initial location following eight moves. The upper right panel shows the final accumulated Dirichlet counts depicting the probability of an aversive outcome at each location. This accumulated evidence—or familiarity with the environment—enables the rat to plan the shortest path to its target after just four exposures. This path is shown as the black dashed line in the left panel. Compare the likelihood mapping with Figure 6. Here, the agent restricted its exploration to those parts of the maze that encompass the path to its goal. The lower panels show an even more restrictive exploration for an unsophisticated rat, which fails to find the shortest path along the white squares. This speaks to the enhanced explorative drive resulting from sophisticated inference.

about beliefs—in virtue of accumulating predictive posterior expectations of expected free energies down plausible paths. In other words, instead of just propagating beliefs about the consequences of successive actions, the scheme simulates belief updating in the future, based on preceding beliefs about the consequences of action. This scheme was illustrated using a simple T-maze problem and a navigation problem that required a deeper search.

In section 1, we noted that active inference may be difficult to scale, although remarkable progress has been made in this direction recently using amortized inference and sampling. For example, Ueltzhöffer (2018) parameterized both the generative model and approximate posterior with function approximators, using evolutionary schemes to minimize variational free energy when gradients were not available. Similarly, Millidge (2019) amortized perception and action by learning a parameterized approximation to expected free energy. Çatal et al. (2019) focused on learning prior preferences, using a learning-from-example approach. Tschantz et al. (2019) extended previous point-estimate models to include full distributions over parameters. This allowed them to apply active inference to continuous control problems (e.g., the mountain car problem, the inverted pendulum task, and a challenging hopper task) and demonstrate an order of magnitude increase in sampling efficiency relative to a strong model-free baseline (Lillcrap et al., 2015). (See Tschantz et al., 2019, for a full discussion and a useful deconstruction of active inference, in relation to things like model-based reinforcement learning; Schrittwieser et al., 2019.)

Note that the navigation example is an instance of planning to learn. As such, it solves the kinds of problems for which reinforcement learning and its variants usually address. In other words, we were able to solve a learning problem from first (i.e., variational) principles without recourse to backward induction or other (belief-free) schemes like Q-learning, SARSA, or successor representations (e.g., Dayan, 1993; Gershman, 2017; Momennejad et al., 2017; Russek, Momennejad, Botvinick, Gershman, & Daw, 2017). This is potentially important because predicating an optimization scheme on inference, as opposed to learning, endows it with a context sensitivity that eludes many learning algorithms (Daw, Gershman, Seymour, Dayan, & Dolan, 2011). In other words, because there are probabilistic representations of time-sensitive hidden states (and implicit uncertainty about those states), behavior is motivated by resolving uncertainty about the context in which an agent is operating. This may be the kind of (Bayesian) mechanics that licenses the notion of competent schemes that can both learn to plan and plan to learn.

The current formulation of active inference does not call on sampling or matrix inversions; the Bayes optimal belief-updating deals with uncertainty in a deterministic fashion. Conceptually, this reflects the difference between the stochastic aspects of random dynamical systems and the deterministic behavior of the accompanying density dynamics, which describe the

probabilistic evolution of those systems (e.g., the Fokker-Planck equation). Because active inference works in belief spaces, that is, on statistical manifolds (Da Costa, Parr, Sengupta, et al., 2020), there is no need for sampling or random searches; the optimal paths are instead evaluated by propagating beliefs or probability distributions into the future to find the path of least variational free energy (Friston, 2013).

In the setting of deep policy searches, this approach has the practical advantage of terminating searches over particular paths when they become implausible. For example, in the navigation example, there were five actions and 64 hidden states, leading to a large number of potential paths ($1.0486 \cdot 10^{10}$ for a planning horizon of four and $1.0737 \cdot 10^{15}$ for a planning horizon of six). However, only a tiny fraction of these paths is actually evaluated—usually several hundred, which takes a few hundred milliseconds on a personal computer. Given reasonably precise beliefs about current states and state transitions, only a small number of paths are eligible for evaluation, which leads us to our final comment on the scalability of active inference.

5.1 Limitations. In one sense, we have addressed scaling through the computational efficiency afforded by belief propagation using a sophisticated scheme. However, we have illustrated this scheme only on rather trivial problems. In principle, one can scale up the dimensionality of state spaces (and outcomes) with a degree of impunity. This follows from the fact that the number of plausible states (and transitions) can be substantially constrained, using the right kind of generative model—one that leverages factorizations and sparsity. For example, the factorization between hidden states and actions used above rests on the implicit assumption that every action is allowed from every state. This is a strong assumption but perfectly apt for many generative models.

One could also call on a related symmetry—namely, a hierarchical separation of temporal scales in deep models, where one Markov decision process is placed on top of another (Friston, Rosch, et al., 2017; George & Hawkins, 2009; Hesp, Smith, et al., 2019; Rikhye et al., 2019). In these models, transitions at the higher level usually unfold at a slower timescale than the level below. This engenders semi-Markovian dependencies that can generate complicated and structured behaviors. In this setting, one could consider hidden states at higher levels that generate the initial and final states of the level below. Policy optimization within each level, using a sophisticated scheme, could then realize the trajectory between the initial states (i.e., empirical priors over initial states) and final states (i.e., priors that determine the cost function and subsequent empirical priors over action).

Finally, it should be noted that in many applications, the states and actions of real-world processes are continuous, which presents a further scaling challenge for discrete state-space models. However, it is possible to

combine sophisticated (discrete) schemes with continuous models, provided one uses the appropriate message passing between the continuous and discrete levels. For example, Friston, Parr, et al. (2017) used a Markov decision process to drive continuous eye movements. Indeed, it would be interesting to revisit simulations of saccadic searches using sophisticated inference, especially in the context of reading.

Appendix: Expected Free Energy

This appendix considers two lemmas that underwrite expected free energy from two complementary perspectives. The first is based on a generative model that combines the principles of optimal Bayesian design (Lindley, 1956) and decision theory (Berger, 2011), while the second is based on a principled account of self-organization (Friston, 2019; Parr et al., 2020). Finally, we consider several corollaries that speak to the notions of active inference (Friston et al., 2015), empowerment (Klyubin, Polani, & Nehaniv, 2005), information bottlenecks (Tishby et al., 1999), self-organization (Friston, 2013), and self-evidencing (Hohwy, 2016). In what follows, $Q(o_\tau, s_\tau, \pi)$ denotes a predictive distribution over future variables and policies, conditioned on initial observations, while $P(o_\tau, s_\tau, \pi)$ denotes a generative model—that is, a marginal distribution over final states and policies. For simplicity, we omit model parameters and assume policies start from the current time point, allowing us to omit the variational free energy from the generalized free energy (since observational evidence is the same for all policies).

A.1 Objective. Our objective is to establish a generalized free energy functional that can be minimized with respect to a posterior over policies, noting that this posterior is necessary to marginalize the joint posterior over hidden states and policies to infer hidden states. To comply with Bayesian decision theory, generalized free energy can be constructed to place an upper bound on Bayesian risk, which corresponds to the divergence between the predictive distribution over outcomes and prior preferences. In other words, Bayesian risk is the expected surprisal or negative log evidence. Confusingly, Bayesian risk and expected risk are two different quantities. The former is the expected surprisal, while the latter is a KL-divergence between predicted and preferred outcomes (or states). To comply with optimal Bayesian design, one can specify priors over policies that lead to states with a precise likelihood mapping to observable outcomes.

Lemma 1 (*Bayes Optimality*). *Generalized free energy¹⁶ is an upper bound on risk, under a generative model whose priors over policies lead to states with precise likelihoods:*

¹⁶Equation A.1 follows from 3.1 when treating $F(\pi)$ and $E(\pi)$ as constants, that is, ignoring past observations and empirical priors over policies.

$$\begin{aligned}
 F[Q(s, \pi)] &= \underbrace{\mathbb{E}_Q[\ln Q(s_\tau, \pi) - \ln P(o_\tau, s_\tau)]}_{\text{Generalised free energy}} \geq \underbrace{D_{KL}[Q(o_\tau)||P(o_\tau)]}_{\text{Risk}} \\
 \log P(\pi) &= \underbrace{\mathbb{E}_{Q(o_\tau, s_\tau|\pi)}[\log P(o_\tau|s_\tau)]}_{\text{Empirical prior}}
 \end{aligned}
 \tag{A.1}$$

Note that $P(\pi)$ is an empirical prior because it depends on the predictive density that depends on past observations. The priors over hidden states and outcomes can be regarded as a target distribution or prior preferences.

Proof. By substituting the empirical prior, equation A.1, into the expression for free energy, we have (noting that policies and outcomes are conditionally independent, given hidden states):

$$\begin{aligned}
 F[Q(s, \pi)] &= \underbrace{\mathbb{E}_Q[D_{KL}[Q(s_\tau|\pi)||P(s_\tau)]]}_{\text{Expected risk (States)}} + \underbrace{D_{KL}[Q(\pi)||P(\pi)]}_{\text{Complexity (Policies)}} \\
 &\geq \underbrace{\mathbb{E}_Q[D_{KL}[Q(s_\tau|\pi)||P(s_\tau)]]}_{\text{Expected risk (States)}} \\
 &= \underbrace{\mathbb{E}_Q[D_{KL}[Q(o_\tau|\pi)||P(o_\tau)]]}_{\text{Expected risk (Outcomes)}} + \underbrace{\mathbb{E}_Q[D_{KL}[Q(s_\tau|o_\tau, \pi)||P(s_\tau|o_\tau)]]}_{\text{Expected evidence bound}} \\
 &\geq \underbrace{\mathbb{E}_Q[D_{KL}[Q(o_\tau|\pi)||P(o_\tau)]]}_{\text{Expected risk (Outcomes)}} \\
 &= \underbrace{D_{KL}[Q(o_\tau)||P(o_\tau)]}_{\text{Risk}} + \underbrace{\mathbb{E}_Q[D_{KL}[Q(o_\tau|\pi)||Q(o_\tau)]]}_{\text{Mutual information}} \\
 &\geq \underbrace{D_{KL}[Q(o_\tau)||P(o_\tau)]}_{\text{Risk}}
 \end{aligned}
 \tag{A.2}$$

These inequalities show that generalized free energy upper bounds the predictive divergence from the marginal likelihood over outcomes (i.e., model evidence). When this bound is minimized, (1) the complexity cost of policies is minimized, enforcing prior beliefs about policies; (2) the predictive posterior over hidden states becomes the posterior under the generative model; and (3) policies and outcomes become independent. This independence follows by construction of the free energy functional and means that final outcomes do not depend on initial conditions, implying a form of steady state (see below). \square

Corollary 1 (*Expected Free Energy*). The free energy can now be minimized with regard to the posterior over policies by expressing free energy in terms of expected free energy:

$$\begin{aligned}
F[Q(s, \pi)] &= \mathbb{E}_{Q(\pi)}[G(\pi) + \ln Q(\pi)] \\
Q(\pi) &= \arg \min_Q F[Q(s, \pi)] \Rightarrow -\ln Q(\pi) = G(\pi) \\
G(\pi) &= \mathbb{E}_{Q(o_\tau, s_\tau | \pi)} \underbrace{[\ln Q(s_\tau | \pi) - \ln P(o_\tau, s_\tau)]}_{\text{Expected free energy}} \\
&= \underbrace{D_{KL}[Q(s_\tau | \pi) || P(s_\tau)]}_{\text{Expected risk}} - \underbrace{\mathbb{E}_{Q(o_\tau, s_\tau | \pi)}[\ln P(o_\tau | s_\tau)]}_{\text{Expected ambiguity}} \tag{A.3}
\end{aligned}$$

This renders free energy $F[Q(s, \pi)] = \mathbb{E}_Q[G(\pi)] - H[Q(\pi)]$ an expected energy minus the entropy of the posterior over policies, in the usual way. Finally, we can express the expected free energy of a policy as a bound on information gain and Bayesian risk:

$$\begin{aligned}
G(\pi) &= \underbrace{\mathbb{E}_Q[D_{KL}[Q(s_\tau | o_\tau, \pi) || P(s_\tau | o_\tau)]]}_{\text{Expected evidence bound}} - \underbrace{\mathbb{E}_Q[\ln P(o_\tau)]}_{\text{Expected log evidence}} \\
&\quad - \underbrace{\mathbb{E}_Q[D_{KL}[Q(s_\tau | o_\tau, \pi) || Q(s_\tau | \pi)]]}_{\text{Expected information gain}} \\
&\geq - \underbrace{\mathbb{E}_Q[D_{KL}[Q(s_\tau | o_\tau, \pi) || Q(s_\tau | \pi)]]}_{\text{Expected information gain}} - \underbrace{\mathbb{E}_Q[\ln P(o_\tau)]}_{\text{Bayesian risk}} \tag{A.4}
\end{aligned}$$

This inequality shows that the free energy of a policy upper bounds a mixture of its expected information gain (Lindley, 1956) and Bayesian risk (Berger, 2011), where Bayesian risk is expected log evidence.

Remark. Here, policies are treated as random variables, which means planning as inference (Attias, 2003; Botvinick & Toussaint, 2012) becomes belief updating under optimal Bayesian design priors (Lindley, 1956; MacKay, 1992). One might ask what licenses these priors above. Although they can be motivated in terms of information gain (see equation A.4), there is a more straightforward motivation that arises as a steady-state solution. We now turn to this complementary perspective that inherits from the Bayesian mechanics described in Friston (2019). Here, we are interested in situations when the predictive distribution attains its steady-state or target distribution.

It may seem odd to predicate optimal behavior on a steady-state distribution. However, the fact that action and its consequences can be expressed probabilistically implies the existence of a (steady-state) joint distribution that does not change over time. In what follows, we use the existence of this steady-state distribution to express the posterior over policies as a functional of the distribution over other variables, given a particular policy. This functional is expected free energy. This represents a deflationary approach to optimality, in the sense that optimal policies are just those that

underwrite a steady state. The question now is, What kind of steady state are we interested in?

We will make a distinction between simple and general steady states in terms of the degeneracy (i.e., many-to-one mapping) of policies to any final state. Simple steady states are characterized by a unique path of least action from some initial observations to a final state. This would be appropriate for describing classical systems, such as a pendulum or planetary bodies. Conversely, a general steady state allows for multiple paths from initial observations to the final state, which means the entropy or uncertainty about which path was actually taken is high. We will be particularly interested in the autonomous behavior of systems whose steady state is maintained by multiple (degenerate) paths or policies. The ensuing distinction can be characterized by a scalar quantity corresponding to the relative entropy or precision of policies and outcomes, conditioned on final states (and initial observations). This scalar $\beta \geq 0$ is not a free parameter; it just characterizes the kind of steady state at hand. Note that in this setup, the notion of optimality is replaced by (or reduces to) the existence of a steady state, which may or may not be simple.

A.2 Objective. We seek distributions over policies that afford steady-state solutions, that is, when the final distribution does not depend on initial observations. Such solutions ensure that on average, stochastic policies lead to a steady-state or target distribution specified by the generative model. These solutions exist in virtue of conditional independencies, where the hidden states provide a Markov blanket (cf. information bottleneck) that separates policies from outcomes. In other words, policies cause final states that cause outcomes. Put simply, policies influence outcomes, but only via hidden states. We will see below that there is a family of such solutions, where the Bayes optimality solution above is a special (canonical) case. In what follows, $Q(o_\tau, s_\tau, \pi) := P(o_\tau, s_\tau, \pi | o_\leq)$ can be read as a posterior distribution, given initial conditions.

Lemma 2 (Nonequilibrium Steady State). *When the surprisal of policies corresponds to a Gibbs free energy $G(\pi, \beta)$, the final distribution attains steady state:*

$$\begin{aligned}
 -\log Q(\pi) &= G(\pi, \beta) \Rightarrow D_{KL}[Q||P] = 0 \\
 G(\pi, \beta) &= \underbrace{D_{KL}[Q(s_\tau|\pi)||P(s_\tau)]}_{\text{Expected risk}} - \underbrace{\mathbb{E}_{Q(o_\tau, s_\tau|\pi)}[\beta \log P(o_\tau|s_\tau)]}_{\text{Expected ambiguity}} \\
 \beta &= \frac{\mathbb{E}_Q[\ln Q(\pi|s_\tau)]}{\mathbb{E}_Q[\ln P(o_\tau|s_\tau)]} = \frac{H(\Pi|S_\tau)}{H(O_\tau|S_\tau)} \\
 P &= P(o_\tau|s_\tau)Q(\pi|s_\tau)P(s_\tau) \\
 Q &= P(o_\tau|s_\tau)Q(s_\tau|\pi)Q(\pi) \tag{A.5}
 \end{aligned}$$

Here, $\beta \geq 0$ characterizes a steady state in terms of the relative precision of policies and final outcomes, given final states. The generative model stipulates steady state, in the sense that distribution over final states (and outcomes) does not depend on initial observations. Here, the generative and predictive distributions simply express the conditional independence between policies and final outcomes, given final states. Note that when $\beta = 1$, Gibbs free energy becomes expected free energy.

Proof. Substituting equation A.5 into the KL divergence between the predictive and generative distributions gives

$$\begin{aligned}
 D_{KL}[Q||P] &= \mathbb{E}_Q \left[\ln \frac{Q(s_\tau|\pi)Q(\pi)}{Q(\pi|s_\tau)P(s_\tau)} \right] \\
 &= \mathbb{E}_Q[\ln Q(\pi) + \ln Q(s_\tau|\pi) - \ln P(s_\tau)] - \mathbb{E}_Q[\ln Q(\pi|s_\tau)] \\
 &= \mathbb{E}_{Q(\pi)}[\ln Q(\pi) + \mathbb{E}_{Q(o_\tau, s_\tau|\pi)}[\ln Q(s_\tau|\pi) - \ln P(s_\tau) \\
 &\quad - \beta \ln P(o_\tau|s_\tau)]] \\
 &= \mathbb{E}_{Q(\pi)}[\ln Q(\pi) + G(\pi, \beta)] \\
 &\Rightarrow \\
 -\ln Q(\pi) = G(\pi, \beta) &\Rightarrow D_{KL}[Q||P] = 0 \Rightarrow \beta = \frac{H(\Pi|S_\tau)}{H(O_\tau|S_\tau)} \quad (\text{A.6})
 \end{aligned}$$

This solution describes a particular kind of steady state, where policies lead to (steady) states with more or less precise likelihoods, depending on the value of β . \square

Remark. At steady state, hidden states (and outcomes) “forget” about initial observations, placing constraints on the distribution over policies that can be expressed in terms of a Gibbs free energy. In the limiting case that $\beta = 0$ (i.e., when $Q(\pi|s)$ tends to a delta function), we obtain a simple steady state where

$$G(\pi, 0) = \mathbb{E}_{Q(o_\tau, s_\tau|\pi)} \left[\ln \frac{Q(s_\tau|\pi)}{P(s_\tau)} \right] = D_{KL}[Q(s_\tau|\pi)||P(s_\tau)] \quad (\text{A.7})$$

This solution corresponds to a standard stochastic control, variously known as KL control or risk-sensitive control (van den Broek et al., 2010). In other words, one picks policies that minimize the divergence between the predictive and target distribution. In different sorts of systems, the relationship between the entropies (β) may differ. As such, different values of this parameter may be appropriate in describing these kinds of system. More generally (i.e., $\beta > 0$), policies are more likely when they lead to states with a precise likelihood mapping. One perspective, on the distinction between simple and general steady states, is in terms of conditional uncertainty about policies. For example, simple (i.e., $\beta = 0$) steady states preclude

uncertainty about which policy led to a final state. This would be appropriate for describing classical systems (that follow a unique path of least action), where it would be possible to infer which policy had been pursued given the initial and final outcomes. Conversely, in general steady-state systems (e.g., mice and men), simply knowing that “you are here” does not tell me “how you got here,” even if I knew where you were this morning. Put another way, there are lots of paths or policies open to systems that attain a general steady state.

The treatment in Friston (2019) effectively turns the steady-state lemma on its head by assuming the steady-state in equation A.5 is stipulatively true—and then characterizes the ensuing self-organization in terms of Bayes optimal policies. In active inference, we are interested in a certain class of systems that self-organize to general steady states: those that move through a large number of probabilistic configurations from their initial state to their final (steady) state. In terms of information geometry, this means that the information distance between any initial and the final (steady) state is large. In the current setting, we could replace information distance (Crooks, 2007; Kim, 2018) by information gain (Lindley, 1956; MacKay, 1992; Still & Precup, 2012). That is, we are interested in systems that attain steady state (i.e., target distributions) with policies associated with a large information gain.¹⁷ Although not pursued here, general steady states with precise likelihood mappings have precise Fisher information matrices and information geometries that distinguish general forms of self-organization from simple forms (Amari, 1998; Ay, 2015; Caticha, 2015; Ikeda, Tanaka, & Amari, 2004; Kim, 2018). This perspective can be unpacked in terms of information theory with the following corollaries, which speak to active inference, empowerment, information bottlenecks, self-organisation, and self-evidencing.

Corollary 2 (*Active Inference*). *If a system attains a general steady state, then by the Bayes optimality lemma, it will appear to behave in a Bayes optimal fashion in terms of both optimal Bayesian design (i.e., exploration) and Bayesian decision theory (i.e., exploitation). Crucially, the loss function defining Bayesian risk is the negative log evidence for the generative model entailed by an agent. In short, systems (i.e., agents) that attain general steady states will look as if they are responding to epistemic affordances (Parr & Friston, 2017).*

Corollary 3 (*Empowerment*). *At its simplest, empowerment (Klyubin et al., 2005) underwrites exploration (i.e., intrinsic motivation) by exploring as many states in the future as possible—and thereby keeping options open. This exploratory*

¹⁷ Note that a divergence such as information gain is not a measure of distance. The information distance (a.k.a. information length) can be regarded as the accumulated divergences along a path on a statistical manifold from the initial location to the final location.

imperative is evinced clearly if we generalize free energy to include β :

$$\begin{aligned} F[Q(s, \pi)] &= \mathbb{E}_Q \left[\ln \frac{Q(s_\tau, \pi)}{P(o_\tau | s_\tau)^\beta P(s_\tau)} \right] = \mathbb{E}_Q \left[\ln \frac{Q(s_\tau | \pi) Q(\pi)}{P(s_\tau) Q(\pi | s_\tau)} \right] \\ &= \underbrace{D_{KL}[Q(s_\tau | \pi) || P(s_\tau)]}_{\text{Risk}} - \underbrace{I(\Pi; S_\tau | o_\leq)}_{\text{Empowerment}} \end{aligned} \quad (\text{A.8})$$

This expresses the free energy of the predictive distribution over final states and policies in terms of risk and empowerment. Minimizing free energy with respect to policies therefore maximizes empowerment—namely, the mutual information between policies and their final states, given initial observations. The epistemic aspect of empowerment can be seen by expressing it in terms of expected ambiguity:

$$\underbrace{I(\Pi; S_\tau | o_\leq)}_{\text{Empowerment}} = \underbrace{H(\Pi | o_\leq)}_{\text{Entropy}} - \underbrace{\mathbb{E}_Q[\beta \ln P(o_\tau | s_\tau)]}_{\text{Expected ambiguity}} \quad (\text{A.9})$$

On this reading, empowerment corresponds to minimizing expected ambiguity while maximizing the entropy of policies—in other words, keeping (policy) options open by avoiding situations from which there is only one escape route. Note that empowerment is a special case of active inference when we can ignore risk (i.e., when all policies are equally risky).

Corollary 4 (Information Bottleneck). The information bottleneck method and related formulations (Bialek, Nemenman, & Tishby, 2001; Still, Sivak, Bell, & Crooks, 2012; Tishby et al., 1999; Tishby & Polani, 2010) can be seen as generalizations of rate distortion theory. According to this view, we can consider hidden states as an information bottleneck (cf. Markov blanket) that plays the role of a compressed representation of past outcomes that best predict future outcomes. Here, we can regard the policies as mapping between initial and final observations via hidden states. The information bottleneck method provides an objective function that can be minimized with respect to the distribution over policies. This (information bottleneck) objective function can be expressed in terms of the expected Gibbs energy as follows:

$$\begin{aligned} \mathbb{E}_{P(o_\leq | \pi)} [G(\pi, \beta)] &= \mathbb{E}_{P(o_\tau, s_\tau, o_\leq | \pi)} \left[\ln \frac{P(s_\tau | o_\leq, \pi)}{P(s_\tau)} + \beta \ln \frac{P(o_\tau)}{P(o_\tau | s_\tau)} - \beta \ln P(o_\tau) \right] \\ &= \underbrace{I(O_\leq; S_\tau | \pi)}_{\text{Information bottleneck}} - \underbrace{\beta I(S_\tau; O_\tau)}_{\text{Bayesian risk}} - \mathbb{E}_P[\beta \ln P(o_\tau)] \end{aligned} \quad (\text{A.10})$$

This means the average Gibbs energy of a policy, over initial observations, combines the information bottleneck objective function and Bayesian risk. Minimizing the first term of the objective function (i.e., the mutual information between initial outcomes and hidden states) plays the role of compression, while maximizing the

second (i.e., the mutual information between hidden states final and outcomes) ensures the information gain that characterizes general steady states. Indeed, when relative precision $\beta = 1$, it is straightforward to show that the information bottleneck is an upper bound on expected information gain:

$$\begin{aligned}
 & \underbrace{I(O_{\leq}; S_{\tau} | \pi) - I(S_{\tau}; O_{\tau})}_{\text{Information bottleneck}} \\
 &= \mathbb{E}_{P(o_{\tau}, s_{\tau}, o_{\leq} | \pi)} [\ln Q(s_{\tau} | \pi) - \ln P(s_{\tau} | o_{\tau})] \\
 &= -\mathbb{E}_{P(o_{\tau} | \pi)} \left[\underbrace{D_{\text{KL}}[Q(s_{\tau} | o_{\tau}, \pi) || Q(s_{\tau} | \pi)]}_{\text{Expected information gain}} + \underbrace{D_{\text{KL}}[Q(s_{\tau} | o_{\tau}, \pi) || P(s_{\tau} | o_{\tau})]}_{\text{Expected evidence bound}} \right] \\
 &\geq -\mathbb{E}_{P(o_{\tau} | \pi)} \left[\underbrace{D_{\text{KL}}[Q(s_{\tau} | o_{\tau}, \pi) || Q(s_{\tau} | \pi)]}_{\text{Expected information gain}} \right] = -I(S_{\tau}; O_{\tau} | O_{\leq}, \pi) \quad (\text{A.11})
 \end{aligned}$$

Because the information bottleneck objective function is an average over initial observations, it cannot be used directly for online (active) planning as inference; however, it can be used to learn fixed outcome-action policies (Hafez-Kolahi & Kasaei, 2019; Tishby & Zaslavsky, 2015). Note that the information bottleneck method is a special case of active inference, when we can ignore Bayesian risk (i.e., when all policies are equally risky).

Corollary 5 (Self-Organization). *The average of expected free energy over policies can be decomposed into risk and conditional entropy:*

$$\begin{aligned}
 \mathbb{E}_{Q(\pi)} [G(\pi)] &= \underbrace{\mathbb{E}_Q [\ln Q(s_{\tau} | \pi) - \ln P(o_{\tau}, s_{\tau})]}_{\text{Expected free energy}} \\
 &= \underbrace{\mathbb{E}_Q [D_{\text{KL}}[Q(s_{\tau} | \pi) || P(s_{\tau})]}_{\text{Expected risk}} + \underbrace{\mathbb{E}_Q [-\ln Q(o_{\tau} | s_{\tau})]}_{\text{Expected ambiguity}} \\
 &= \underbrace{\mathbb{E}_Q [D_{\text{KL}}[Q(s_{\tau} | \pi) || P(s_{\tau})]}_{\text{Expected risk}} + \underbrace{H(O_{\tau} | S_{\tau}, o_{\leq})}_{\text{Conditional entropy}} \geq 0 \quad (\text{A.12})
 \end{aligned}$$

This decomposition means that if the expected free energy of policies is small on average, the predictive distribution over hidden states will converge to the prior or preferred distribution, while uncertainty about consequent outcomes will be small. In the limit, the predictive distribution over hidden states becomes the prior distribution, with no uncertainty about outcomes. This can be read as the limiting case of self-organization to prior beliefs.

Corollary 6 (Self-Evidencing). *The average of expected free energy over policies furnishes an upper bound on the (negative) expected log evidence of outcomes*

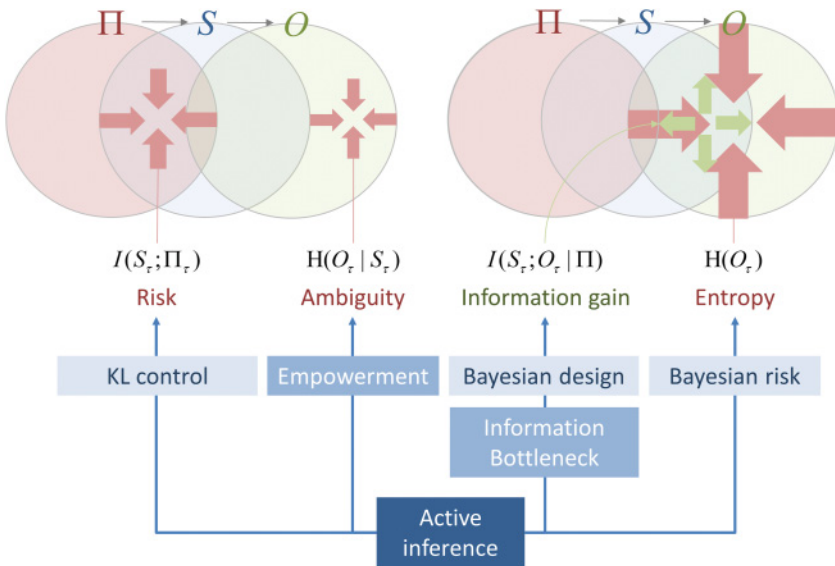


Figure 8: Active inference and other schemes. This schematic summarizes the various imperatives implied by minimizing a free energy functional of posterior beliefs about policies, ensuing states, and subsequent outcomes. The information diagrams in the upper panels represent the entropy of the three variables, where intersections correspond to shared information or mutual information. A conditional entropy corresponds to an area that precludes the variable on which the entropy is conditioned. Note that there is no overlap between policies and outcomes that is outside hidden states. This is because hidden states form a Markov blanket (i.e., information bottleneck) between policies and outcomes. Two complementary formulations of minimizing expected free energy are shown on the right (in terms of risk and ambiguity) and left (in terms of information gain and entropy), respectively. Both will tend to increase the overlap or mutual information between hidden states and outputs while minimizing entropy or Bayesian risk. In these diagrams, we have assumed steady state, such that risk becomes the mutual information between policies and hidden states. For simplicity, we have omitted dependencies on initial observations. The various schemes or formulations considered in the text are shown at the bottom. These demonstrate that Bayesian decision theory (i.e., KL control and Bayesian risk) and optimal Bayesian design figure as complementary imperatives.

and the mutual information between these outcomes and their causes (i.e., hidden states):

$$\mathbb{E}_{Q(\pi)}[G(\pi)] = \underbrace{\mathbb{E}_Q[\ln Q(s_\tau|\pi) - \ln P(o_\tau, s_\tau)]}_{\text{Expected free energy}}$$

$$\begin{aligned}
&= - \underbrace{\mathbb{E}_{Q(o_\tau, \pi)} [D_{KL}[Q(s_\tau | o_\tau, \pi) || Q(s_\tau | \pi)]]}_{\text{Expected information gain}} - \underbrace{\mathbb{E}_{Q(o_\tau)} [\ln P(o_\tau)]}_{\text{Expected log evidence}} \\
&\quad + \underbrace{\mathbb{E}_{Q(o_\tau, \pi)} [D_{KL}[Q(s_\tau | o_\tau, \pi) || P(s_\tau | o_\tau)]]}_{\text{Expected evidence bound}} \\
&\geq - \underbrace{I(S_\tau, O_\tau | \Pi, o_\leq)}_{\text{Mutual information}} - \underbrace{\mathbb{E}_{Q(o_\tau)} [\ln P(o_\tau)]}_{\text{Expected log evidence}} \tag{A.13}
\end{aligned}$$

This decomposition means that if the expected free energy of policies is, on average, small, the expected log evidence and the mutual information between predicted states and the outcomes they generate will be large. In the limit, expected log evidence is maximized, with no uncertainty about outcomes, given hidden states. This can be read as the limiting case of self-evidencing with unambiguous outcomes.

It can sometimes be difficult to see the relationships between the various conditional entropy and mutual information terms that constitute the free energy functional. Figure 8 tries to clarify these relationships using information diagrams. This schematic highlights the complementary decompositions of expected free energy in terms of risk and ambiguity—and information gain and entropy. These decompositions are summarized in terms of the imperative to minimize various segments of the information diagrams. Figure 8 then highlights the particular components that figure in special cases, such as an optimal Bayesian decisions and design.

Software Note

Although the generative model changes from application to application, the belief updates described in this letter are generic and can be implemented using standard routines (here, `spm_MDP_VB_XX.m`). These routines are available as Matlab code in the SPM academic software: <http://www.fil.ion.ucl.ac.uk/spm/>. The simulations in this letter can be reproduced (and customized) via a graphical user interface by typing `>> DEM` and selecting the appropriate **(T-maze or Navigation) demo**.

Acknowledgments

K.J.F. was funded by the Wellcome Trust (088130/Z/09/Z). L.D. is supported by the Fonds National de la Recherche, Luxembourg (13568875). C.H. was funded by a Research Talent Grant (406.18.535) of the Netherlands Organisation for Scientific Research. We have no disclosures or conflicts of interest.

References

- Amari, S. (1998). Natural gradient works efficiently in learning. *Neural Computation*, 10(2), 251–276. doi:10.1162/089976698300017746
- Åström, K. J. (1965). Optimal control of Markov processes with incomplete state information. *Journal of Mathematical Analysis and Applications*, 10(1), 174–205.
- Attias, H. (2003). Planning by probabilistic inference. In *Proceedings of the 9th Int. Workshop on Artificial Intelligence and Statistics*. New York: ACM.
- Ay, N. (2015). Information geometry on complexity and stochastic interaction. *Entropy*, 17(4), 2432.
- Barlow, H. (1961). Possible principles underlying the transformations of sensory messages. In W. Rosenblith (Ed.), *Sensory communication* (pp. 217–234). Cambridge, MA: MIT Press.
- Barlow, H. B. (1974). Inductive inference, coding, perception, and language. *Perception*, 3, 123–134.
- Barto, A., Mirolli, M., & Baldassarre, G. (2013). Novelty or surprise? *Frontiers in Psychology*, 4. doi:10.3389/fpsyg.2013.00907
- Beal, M. J. (2003). *Variational algorithms for approximate Bayesian inference*. PhD diss., University College London.
- Bellman, R. (1952). On the theory of dynamic programming. In *Proc. Natl. Acad. Sci. USA*, 38, 716–719.
- Berger, J. O. (2011). *Statistical decision theory and Bayesian analysis*. New York: Springer.
- Bialek, W., Nemenman, I., & Tishby, N. (2001). Predictability, complexity, and learning. *Neural Comput.*, 13(11), 2409–2463.
- Botvinick, M., & Toussaint, M. (2012). Planning as inference. *Trends Cogn. Sci.*, 16(10), 485–488.
- Camerer, C. F., Ho, T.-H., & Chong, J.-K. (2004). A cognitive hierarchy model of games. *Quarterly Journal of Economics*, 119(3), 861–898. doi:10.1162/0033553041502225
- Catal, O., Nauta, J., Verbelen, T., Simoens, P., & Dhoedt, B. (2019). *Bayesian policy selection using active inference*. <https://arxiv.org/pdf/1904.08149.pdf>
- Çatal, O., Verbelen, T., Nauta, J., Boom, C. D., & Dhoedt, B. (2020). Learning perception and planning with deep active inference. In *Proceedings of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing*. Piscataway, NJ: IEEE.
- Çatal, O., Wauthier, S., Verbelen, T., De Boom, C., & Dhoedt, B. (2020). *Deep active inference for autonomous robot navigation*. arXiv:2003.03220.
- Caticha, A. (2015). The basics of information geometry. In *Proceedings of the AIP Conference Proceedings* (pp. 15–26). College Park, MD: American Institute of Physics. doi:10.1063/1.4905960
- Cohen, J. D., McClure, S. M., & Yu, A. J. (2007). Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.*, 362(1481), 933–942.
- Costa-Gomes, M., Crawford, V. P., & Broseta, B. (2001). Cognition and behavior in normal-form games: An experimental study. *Econometrica*, 69(5), 1193–1235. doi:10.1111/1468-0262.00239

- Crooks, G. E. (2007). Measuring thermodynamic length. *Phys. Rev. Lett.*, 99(10), 100602. doi:10.1103/PhysRevLett.99.100602
- Da Costa, L., Parr, T., Sajid, N., Veselic, S., Neacsu, V., & Friston, K. (2020). *Active inference on discrete state-spaces: A synthesis*. arXiv:2001.07203.
- Da Costa, L., Parr, T., Sengupta, B., & Friston, K. (2020). *Natural selection finds natural gradient*. arXiv:2001.08028.
- Da Costa, L., Sajid, N., Parr, T., Friston, K. J., & Smith, R. (2020). *The relationship between dynamic programming and active inference: The discrete, finite-horizon case*. arXiv:2009.08111.
- Dauwels, J. (2007). *On variational message passing on factor graphs*. In *Proceedings of the 2007 IEEE International Symposium on Information Theory*. Piscataway, NJ: IEEE.
- Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron*, 69(6), 1204–1215.
- Dayan, P. (1993). Improving generalization for temporal difference learning: The successor representation. *Neural Comput.*, 5(4), 613–624. doi:10.1162/neco.1993.5.4.613
- Dayan, P., Hinton, G. E., Neal, R. M., & Zemel, R. S. (1995). The Helmholtz machine. *Neural Comput.*, 7(5), 889–904.
- Devaine, M., Hollard, G., & Daunizeau, J. (2014). Theory of mind: Did evolution fool us? *PLOS One*, 9(2), e87619. doi:10.1371/journal.pone.0087619
- Doya, K. (1999). What are the computations of the cerebellum, the basal ganglia and the cerebral cortex? *Neural Netw.*, 12(7–8), 961–974. doi:10.1016/s0893-6080(99)00046-5
- Duff, M. O. (2002). *Optimal learning: Computational procedures for Bayes-adaptive Markov decision processes*. PhD diss., University of Massachusetts.
- Fleming, W. H., & Sheu, S. J. (2002). Risk-sensitive control and an optimal investment model II. *Ann. Appl. Probab.*, 12(2), 730–767. doi:10.1214/aoap/1026915623
- Friston, K. (2013). Life as we know it. *J. R. Soc. Interface*, 10(86), 20130475.
- Friston, K. (2019). *A free energy principle for a particular physics*. arXiv:1906.10184.
- Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., O'Doherty, J., & Pezzulo, G. (2016). Active inference and learning. *Neurosci. Biobehav. Rev.*, 68, 862–879. doi:10.1016/j.neubiorev.2016.06.022
- Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., & Pezzulo, G. (2017). Active inference: A process theory. *Neural Comput.*, 29(1), 1–49. doi:10.1162/NECO_a_00912
- Friston, K. J., Lin, M., Frith, C. D., Pezzulo, G., Hobson, J. A., & Ondobaka, S. (2017). Active inference, curiosity and insight. *Neural Comput.*, 29(10), 2633–2683. doi:10.1162/neco_a_00999
- Friston, K. J., Parr, T., & de Vries, B. (2017). The graphical brain: Belief propagation and active inference. *Netw. Neurosci.*, 1(4), 381–414. doi:10.1162/NETN_a_00018
- Friston, K., Rigoli, F., Ognibene, D., Mathys, C., Fitzgerald, T., & Pezzulo, G. (2015). Active inference and epistemic value. *Cogn. Neurosci.*, 6(4), 187–214. doi:10.1080/17588928.2015.1020053
- Friston, K. J., Rosch, R., Parr, T., Price, C., & Bowman, H. (2017). Deep temporal models and active inference. *Neurosci. Biobehav. Rev.*, 77, 388–402. doi:10.1016/j.neubiorev.2017.04.009

- George, D., & Hawkins, J. (2009). Towards a mathematical theory of cortical micro-circuits. *PLOS Comput. Biol.*, *5*(10), e1000532. doi:10.1371/journal.pcbi.1000532
- Gershman, S. J. (2017). Predicting the past, remembering the future. *Curr. Opin. Behav. Sci.*, *17*, 7–13. doi:10.1016/j.cobeha.2017.05.025
- Ghahramani, Z., & Jordan, M. I. (1997). Factorial hidden Markov models. *Machine Learning*, *29*(2–3), 245–273. doi:10.1023/a:1007425814087
- Ghavamzadeh, M., Mannor, S., Pineau, J., & Tamar, A. (2016). *Bayesian reinforcement learning: A survey*. arXiv:1609.04436.
- Hafez-Kolahi, H., & Kasaei, S. (2019). *Information bottleneck and its applications in deep learning*. CoRR, abs/1904.03743.
- Hesp, C., Ramstead, M., Constant, A., Badcockyy, P., Kirchhoff, M., & Friston, K. (2019). A multi-scale view of the emergent complexity of life: A free-energy proposal. In G. Georgiev, J. Smart, C. Flores Martinez, & M. Price (Eds.) *Evolution, development and complexity* (pp. 195–227). Cham: Springer.
- Hesp, C., Smith, R., Parr, T., Allen, M., Friston, K., & Ramstead, M. (2019). *Deeply felt affect: The emergence of valence in deep active inference*. PsyArXiv. doi:10.31234/osf.io/62pfd
- Hohwy, J. (2016). The self-evidencing brain. *Nous*, *50*(2), 259–285. doi:10.1111/nous.12062
- Howard, R. (1966). Information value theory. *IEEE Transactions on Systems, Science and Cybernetics*, *SSC-2*(1), 22–26.
- Ikeda, S., Tanaka, T., & Amari, S.-I. (2004). Stochastic reasoning, free energy, and information geometry. *Neural Computation*, *16*, 1779–1810. doi:10.1162/0899766041336477
- Itti, L., & Baldi, P. (2009). Bayesian surprise attracts human attention. *Vision Res.*, *49*(10), 1295–1306.
- Jaynes, E. T. (1957). Information theory and statistical mechanics. *Physical Review Series II*, *106*(4), 620–630.
- Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*(2), 263–291.
- Kaplan, R., & Friston, K. J. (2018). Planning and navigation as active inference. *Biol. Cybern.*, *112*(4), 323–343. doi:10.1007/s00422-018-0753-2
- Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773–795. doi:10.1080/01621459.1995.10476572
- Kauder, E. (1953). Genesis of the marginal utility theory: From Aristotle to the end of the eighteenth century. *Economic Journal*, *63*(251), 638–650.
- Keramati, M., Smittenaar, P., Dolan, R. J., & Dayan, P. (2016). Adaptive integration of habits into depth-limited planning defines a habitual-goal-directed spectrum. In *Proc. Natl. Acad. Sci. U.S.A.*, *113*(45), 12868–12873. doi:10.1073/pnas.1609094113
- Kim, E.-j. (2018). Investigating information geometry in classical and quantum systems through information length. *Entropy*, *20*(8), 574. doi:10.3390/e20080574
- Klyubin, A. S., Polani, D., & Nehaniv, C. L. (2005). Empowerment: A universal agent-centric measure of control. In *Proceedings of the IEEE Congress on Evolutionary Computation* (1:128–135). Piscataway, NJ: IEEE.
- Lee, J. J., & Keramati, M. (2017). Flexibility to contingency changes distinguishes habitual and goal-directed strategies in humans. *PLOS Comput. Biol.*, *13*(9). e1005753. doi:10.1371/journal.pcbi.1005753

- Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., . . . Wierstra, D. (2015). *Continuous control with deep reinforcement learning*. arXiv:1509.02971.
- Lindley, D. V. (1956). On a measure of the information provided by an experiment. *Ann. Math. Statist.*, 27(4), 986–1005. doi:10.1214/aoms/1177728069
- Linsker, R. (1990). Perceptual neural organization: Some approaches based on network models and information theory. *Annu. Rev. Neurosci.*, 13, 257–281.
- MacKay, D. J. C. (1992). Information-based objective functions for active data selection. *Neural Computation*, 4(4), 590–604. doi:10.1162/neco.1992.4.4.590
- MacKay, D. J. C. (2003). *Information theory, inference and learning algorithms*. Cambridge: Cambridge University Press.
- Millidge, B. (2019). *Deep active inference as variational policy gradients*. arXiv:1907.03876.
- Momennejad, I., Russek, E. M., Cheong, J. H., Botvinick, M. M., Daw, N. D., & Gershman, S. J. (2017). The successor representation in human reinforcement learning. *Nature Human Behavior*, 1(9), 680–692. doi:10.1038/s41562-017-0180-8
- Optican, L., & Richmond, B. J. (1987). Temporal encoding of two-dimensional patterns by single units in primate inferior cortex. II: Information theoretic analysis. *J. Neurophysiol.*, 57, 132–146.
- Ortega, P. A., & Braun, D. A. (2010). A minimum relative entropy principle for learning and acting. *Journal of Artificial Intelligence Research*, 38, 475–511.
- Osband, I., Van Roy, B., Russo, D. J., & Wen, Z. (2019). Deep exploration via randomized value functions. *Journal of Machine Learning Research*, 20(124), 1–62.
- Oudeyer, P.-Y., & Kaplan, F. (2007). What is intrinsic motivation? A typology of computational approaches. *Frontiers in Neurorobotics*, 1, 6.
- Parr, T., Da Costa, L., & Friston, K. (2020). Markov blankets, information geometry and stochastic thermodynamics. *Philosophical Transactions of the Royal Society A*, 378(2164).
- Parr, T., & Friston, K. J. (2017). Working memory, attention, and salience in active inference. *Sci. Rep.*, 7(1), 14678. doi:10.1038/s41598-017-15249-0
- Parr, T., & Friston, K. J. (2019a). Attention or salience? *Current Opinion in Psychology*, 29, 1–5. doi: <https://doi.org/10.1016/j.copsyc.2018.10.006>
- Parr, T., & Friston, K. J. (2019b). Generalized free energy and active inference. *Biol. Cybern.*, 113(5–6), 495–513. doi:10.1007/s00422-019-00805-w
- Parr, T., Markovic, D., Kiebel, S. J., & Friston, K. J. (2019). Neuronal message passing using mean-field, Bethe, and marginal approximations. *Sci. Rep.*, 9(1), 1889. doi:10.1038/s41598-018-38246-3
- Ramnani, N. (2014). Automatic and controlled processing in the corticocerebellar system. *Prog. Brain Res.*, 210, 255–285. doi:10.1016/b978-0-444-63356-9.00010-8
- Rikhye, R. V., Guntupalli, J. S., Gothoskar, N., Lázaro-Gredilla, M., & George, D. (2019). *Memorize-generalize: An online algorithm for learning higher-order sequential structure with cloned hidden Markov models*. bioRxiv:764456. doi:10.1101/764456
- Ross, S., Chaib-draa, B., & Pineau, J. (2008). Bayes-adaptive POMDPs. In J. C. Platt, D. Koller, Y. Singer, & S. T. Roweis (Eds.), *Advances in neural information processing systems*, 20. Cambridge, MA: MIT Press.
- Russek, E. M., Momennejad, I., Botvinick, M. M., Gershman, S. J., & Daw, N. D. (2017). Predictive representations can link model-based reinforcement learning to model-free mechanisms. *PLOS Comput. Biol.*, 13(9), e1005768. doi:10.1371/journal.pcbi.1005768

- Ryan, R., & Deci, E. (1985). *Intrinsic motivation and self-determination in human behavior*. New York: Plenum.
- Schmidhuber, J. (1991). Curious model-building control systems. In *Proceedings of the 1991 IEEE International Joint Conference on Neural Network* (pp. 1458–1463). Piscataway, NJ: IEEE. doi:10.1109/IJCNN.1991.170605
- Schmidhuber, J. (2006). Developmental robotics, optimal artificial curiosity, creativity, music, and the fine arts. *Connection Science*, 18(2), 173–187. doi:10.1080/09540090600768658
- Schmidhuber, J. (2010). Formal theory of creativity, fun, and intrinsic motivation (1990–2010). *IEEE Transactions on Autonomous Mental Development*, 2(3), 230–247. doi:10.1109/tamd.2010.2056368
- Schrittwieser, J., Antonoglou, I., Hubert, T., Simonyan, K., Sifre, L., Schmitt, S., . . . Silver, D. (2019). *Mastering Atari, go, chess and shogi by planning with a learned model*. arXiv:1911.08265.
- Schwartenbeck, P., Fitzgerald, T., Dolan, R. J., & Friston, K. (2013). Exploration, novelty, surprise, and free energy minimization. *Front. Psychol.*, 4, 710. doi:10.3389/fpsyg.2013.00710
- Schwartenbeck, P., Passerker, J., Hauser, T. U., FitzGerald, T. H. B., Kronbichler, M., & Friston, K. J. (2019). Computational mechanisms of curiosity and goal-directed exploration. *Elife*, 8, e41703. doi:10.7554/eLife.41703
- Sengupta, B., & Friston, K. (2018). *How robust are deep neural networks?* arXiv:1804.11313.
- Solway, A., & Botvinick, M. M. (2015). Evidence integration in model-based tree search. In *Proc. Natl. Acad. Sci. U.S.A.*, 112(37), 11708–11713. doi:10.1073/pnas.1505483112
- Still, S., & Precup, D. (2012). An information-theoretic approach to curiosity-driven reinforcement learning. *Theory Biosci.*, 131(3), 139–148. doi:10.1007/s12064-011-0142-z
- Still, S., Sivak, D. A., Bell, A. J., & Crooks, G. E. (2012). Thermodynamics of prediction. *Phys. Rev. Lett.*, 109(12), 120604. doi:10.1103/PhysRevLett.109.120604
- Suh, S., Chae, D. H., Kang, H. G., & Choi, S. (2016). Echo-state conditional variational autoencoder for anomaly detection. In *Proceedings of the 2016 International Joint Conference on Neural Networks* (pp. 1015–1022). Berlin: Springer.
- Sun, Y., Gomez, F., & Schmidhuber, J. (2011). Planning to be surprised: Optimal Bayesian exploration in dynamic environments. In J. Schmidhuber, K. R. Thórisson, & M. Looks (Eds.), *Proceedings of the Artificial General Intelligence 4th International Conference*, (pp. 41–51). Berlin: Springer.
- Sutton, R. S., & Barto, A. G. (1981). Toward a modern theory of adaptive networks: Expectation and prediction. *Psychol. Rev.*, 88(2), 135–170.
- Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. Cambridge, MA: MIT Press.
- Thompson, W. R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3–4), 285–294. doi:10.1093/biomet/25.3-4.285
- Tishby, N., Pereira, F. C., & Bialek, W. (1999). The information bottleneck method. In *Proceedings of the 37th Annual Allerton Conference on Communication, Control, and Computing* (pp. 368–377). Champaign: University of Illinois.

- Tishby, N., & Polani, D. (2010). Information theory of decisions and actions. In V. Cutsuridis, A. Hussain, & J. Taylor (Eds.), *Perception-reason-action cycle: Models, algorithms and systems*. Berlin: Springer.
- Tishby, N., & Zaslavsky, N. (2015). *Deep learning and the information bottleneck principle*. arXiv:1503.02406
- Todorov, E. (2008). General duality between optimal control and estimation. In *Proceedings of the IEEE Conference on Decision and Control*. Piscataway, NJ: IEEE.
- Toussaint, M., & Storkey, A. (2006). Probabilistic inference for solving discrete and continuous state Markov decision processes. In *Proceedings of the 23rd Int. Conf. on Machine Learning*. New York: ACM.
- Tribus, M. (1961). *Thermodynamics and thermostatics: An introduction to energy, information and states of matter, with engineering applications*. New York: Van Nostrand.
- Tschantz, A., Baltieri, M., Seth, A. K., & Buckley, C. L. (2019). *Scaling active inference*. arXiv:1911.10601.
- Ueltzhöffer, K. (2018). Deep active inference. *Biol. Cybern.*, 112(6), 547–573. doi:10.1007/s00422-018-0785-7
- van den Broek, J. L., Wiegierinck, W. A. J. J., & Kappen, H. J. (2010). Risk-sensitive path integral control. *Uncertainty in Artificial Intelligence*, 6, 1–8.
- Von Neumann, J., & Morgenstern, O. (1944). *Theory of games and economic behavior*. Princeton: Princeton University Press.
- Winn, J., & Bishop, C. M. (2005). Variational message passing. *Journal of Machine Learning Research*, 6, 661–694.
- Yedidia, J. S., Freeman, W. T., & Weiss, Y. (2005). Constructing free energy approximations and generalized belief propagation algorithms. *IEEE Transactions on Information Theory*, 51, 2282–2312.

Received June 8, 2020; accepted September 18, 2020.